PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 1/21, 5/10, 15/12, 15/64, 15/66, 15/70, 15/79, C07K 14/705

A1

(11) International Publication Number: **WO 98/20110**

(43) International Publication Date: 14 May 1998 (14.05.98)

(21) International Application Number: PCT/US96/17549

(22) International Filing Date: 1 November 1996 (01.11.96)

(71) Applicants *(for all designated States except US)*: HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850 (US). THE UNIVERSITY OF AUCKLAND [NZ/NZ]; 85 Park Road, Grafton, Auckland 1001 (NZ).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: NI, Jian [CN/US]; 5502 Manorfield Road, Rockville, MD 20853 (US). GREENE, John, M. [US/US]; 872 Diamond Drive, Gaithersburg, MD 20878 (US). KRISSANSEN, Geoffrey, W. [NZ/NZ]; 157B Grand Drive, St. Johns, Auckland 1001 (NZ). LEUNG, Euphemia, Yee, Fun [US/NZ]; 10 Martin Avenue, Remuera, Auckland 1001 (NZ). RUBEN, Steven, M. [US/US]; 18528 Heritage Hills Drive, Olney, MD 20832 (US).

(74) Agents: GOLDSTEIN, Jorge, A. et al.; Sterne, Kessler, Goldstein & Fox P.L.L.C., Suite 600, 1100 New York Avenue, N.W., Washington, DC 20005–3934 (US).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
With an indication in relation to a deposited microorganism furnished under Rule 13bis separately from the description.
Date of receipt by the International Bureau:
10 October 1996 (10.10.1996)

(54) Title: HUMAN MUCOSAL ADDRESSIN CELL ADHESION MOLECULE–1 (MAdCAM–1) AND SPLICE VARIANTS THEREOF

(57) Abstract

The present invention relates to novel MAdCAM–1 proteins designated herein as MAdCAM–1(a–e), which are cell adhesion molecules. In particular, isolated nucleic acid molecules are provided encoding the human MAdCAM–1(a–e) proteins. MAdCAM–1(a–e) polypeptides are also provided as are vectors, host cells and recombinant methods for producing the same. The invention further relates to screening methods for identifying agonists and antagonists of MAdCAM–1(a–e) activity. Also provided are diagnostic methods for detecting cancer or a pathological inflammatory condition, and therapeutic methods for treating an individual in need of a reduction in the activity of any of MAdCAM–1(a–e). In another aspect, the invention provides isolated genomic DNA molecules comprising the 5 exons which comprise the genes which encode any of MAdCAM–1(a–e), as well as the 5' flanking region which includes the promoter for these genes. In another aspect, the invention relates to a method of screening compounds for the ability to regulate expression of any of MAdCAM–1(a–e) from their promoter. The invention also relates to a method of selectively expressing genes on gut endothelia.

Number 97758 on October 10, 1996. The nucleotide sequence determined by sequencing portions of the deposited genomic DNA, which is shown in FIG. 6, includes the sequence of the 5' flanking region, given in SEQ ID NO:33, as well as the sequences of exons 1-5, given in SEQ ID NOS:34-38, respectively.

The invention further provides isolated MAdCAM-1 polypeptides (MAdCAM-1(a-e)) having an amino acid sequence encoded by a polynucleotide described herein.

The present invention also provides a screening method for identifying compounds capable of enhancing or inhibiting a cellular response induced by any of the MAdCAM-1 polypeptides (designated MAdCAM-1(a-e)), which involves contacting cells which express the desired MAdCAM-1 polypeptides with the candidate compound, assaying a cellular response, and comparing the cellular response to a standard cellular response, the standard being assayed when contact is made in absence of the candidate compound; whereby, an increased cellular response over the standard indicates that the compound is an agonist and a decreased cellular response over the standard indicates that the compound is an antagonist.

The invention also provides a diagnostic method useful during diagnosis of an inflammatory disorder.

An additional aspect of the invention is related to a method for treating an individual in need of a decreased level of MAdCAM-1(a-e) activity in the body comprising administering to such an individual a composition comprising a therapeutically effective amount of an antagonist of MAdCAM-1(a-e)-mediated adhesion. Preferred antagonists for use in the present invention are MAdCAM-1(a-e)-specific antibodies, as well as soluble forms of MAdCAM-1(a-e).

As the invention also includes isolated genomic DNA molecules comprising the 5' flanking region of MAdCAM-1(a-e), including the promoter for these genes, yet another aspect of the invention is related to a method for identifying compounds capable of enhancing or inhibiting expression of any of MAdCAM-1(a-e).

Because MAdCAM-1 is selectively expressed on HEV and on lamina propria venules, the promoter can also be used to selectively target therapeutic genes to the gut endothelia.

Brief Description of the Figures

FIGS. 1A and 1B show the nucleotide (SEQ ID NO:1) and deduced amino acid (SEQ ID NO:2) sequences of MAdCAM-1(a). The protein has a leader sequence of about 17 amino acid residues (first underlined region), followed by an extracellular domain. The second underlined region corresponds to the transmembrane domain, and is followed by the intracellular domain. The predicted amino acid sequence of the mature MAdCAM-1(a) protein (which lacks the leader sequence) is also shown in FIG. 1 (SEQ ID NO:2).

FIGS. 2A and 2B show the nucleotide (SEQ ID NO:3) and deduced amino acid (SEQ ID NO:4) sequences of MAdCAM-1(b). The protein has a leader sequence of about 17 amino acid residues (first underlined region), followed by an extracellular domain. The second underlined region corresponds to the transmembrane domain, and is followed by the intracellular domain. The predicted amino acid sequence of the mature MAdCAM-1(b) protein (which lacks the leader sequence) is also shown in FIG. 2 (SEQ ID NO:4).

FIG. 3 shows the nucleotide (SEQ ID NO:5) and deduced amino acid (SEQ ID NO:6) sequences of MAdCAM-1(c). The protein has a leader sequence of about 17 amino acid residues (first underlined region), followed by an extracellular domain. The second underlined region corresponds to the transmembrane domain, and is followed by the intracellular domain. The predicted amino acid sequence of the mature MAdCAM-1(b) protein (which lacks the leader sequence) is also shown in FIG. 3 (SEQ ID NO:6).

FIGS. 4A and 4B show the nucleotide (SEQ ID NO:7) and deduced amino acid (SEQ ID NO:8) sequences of MAdCAM-1(d). The protein has a leader sequence of about 17 amino acid residues (first underlined region), followed by an extracellular domain. The second underlined region corresponds to the transmembrane domain, and is followed by the intracellular domain. The predicted amino acid sequence of the mature MAdCAM-1(d) protein (which lacks the leader sequence) is also shown in FIG. 4 (SEQ ID NO:8).

FIGS. 5A and 5B show the nucleotide (SEQ ID NO:9) and deduced amino acid (SEQ ID NO:10) sequences of MAdCAM-1(e). The protein has a leader sequence of about 17 amino acid residues (first underlined region), followed by an extracellular domain. The second underlined region corresponds to the transmembrane domain, and is followed by the intracellular domain. The predicted amino acid sequence of the mature MAdCAM-1(e) protein (which lacks the leader sequence) is also shown in FIG. 5 (SEQ ID NO:10).

FIGS. 6A and 6B show the nucleotide sequence of genomic DNA encoding the region 5' to the gene encoding MAdCAM-1 (SEQ ID NO:33). Also shown are exons 1-5 (SEQ ID NOS:34-38, respectively), which comprise the genes which encode any of MAdCAM-1(a-e). Lower case letters represent intron sequence.

FIGS. 7A and 7B show the regions of similarity between the predicted amino acid sequences of the human MAdCAM-1(a-e) proteins (SEQ ID NOS:2, 4, 6, 8, 10, respectively), mouse MAdCAM-1 (SEQ ID NO:46), and the predicted amino acid sequence of human MAdCAM-1 from Shyjan *et al., J. Immunol.* *156*(8):2851-2857 (1996) (SEQ ID NO:47).

FIG. 8 shows an analysis of the MAdCAM-1(a) amino acid sequence. Alpha, beta, turn and coil regions; hydrophilicity and hydrophobicity;

amphipathic regions; flexible regions; antigenic index and surface probability are shown. In the "Antigenic Index - Jameson-Wolf" graph, amino acid residues 52-80, 164-296 and 228-321 in FIG. 1 correspond to the shown highly antigenic regions of the MAdCAM-1(a) protein.

FIG. 9A shows the isolation of MAdCAM-1(a) cDNA. MAdCAM-1(a) cDNAs were initially identified as expressed sequence tags (ESTs), clones HEBBC23X and Y, in an EST database created from an early stage human brain cDNA library. The insert of clone HEBBC23Y was subsequently used to isolate clone MAD-C1 from a human cosmid library. Complementary DNA encoding the 5'-end of human MAdCAM-1(a) was obtained by PCR using PCR primers designed from HEBBC23X and MAD-C1, yielding PCR clone PCR1-5'. The upper FIG. illustrates a partial restriction map of the composite cDNA sequence derived from the overlapping partial clones. The boxed region denotes the open reading frame; and the restriction enzyme sites are marked with vertical lines. FIG. 9B shows nucleotide and deduced amino acid sequence of human MAdCAM-1(a) (SEQ ID NOS:1 and 2). The numbers in the right-hand margin show nucleotide and amino acid positions, respectively. The initiation methionine has been assigned to position 1 by comparison with the mouse MAdCAM-1(a) sequence. The putative signal peptide and transmembrane domains are underlined. The major (residues 226 to 273) mucin domain is boxed, and the minor mucin (residues 278 to 311) domain is italicized, and cysteines expected to form disulphide bonds in the two immunoglobulin domains are circled. A potential polyadenylation signal site is overlined.

FIGS. 10A and 10B show a comparison of the major mucin domain of human MAdCAM-1(a) with the imperfect repeats of the mucin domain of the intestinal mucin MUC-2. In FIG. 10A, the six octomer repeats comprising the major mucin domain of MAdCAM-1(a) have been aligned (SEQ ID NOS:49, 50, 50, 51, 51, and 52, respectively), and shared residues are indicated by bold type.

In FIG. 10B, the six repeats of the MAdCAM-1(a) major mucin domain (SEQ ID NO:53) and MUC-2 (SEQ ID NO 55) are optimally aligned (comparison is SEQ ID NO:54). Identical amino acids are indicated, and conservative substitutions are denoted (+). Numbers refer to amino acid residues.

FIGS. 11A and 11B show an identification of MAdCAM-1 splice variants (SEQ ID NO:2). In FIG. 11A, partial sequences of MAdCAM-1 splice variants encoding the second Ig domain and the major mucin domain or parts thereof have been aligned. HEBBC23Y, which is missing 3 mucin repeats, was identified as an EST. Sequences 3, 5 and 7 are missing a major portion of the second Ig domain and 3 to 6 mucin repeats were isolated as PCR products following amplification from fetal brain RNA. In FIG. 11B, sequences of acceptor and donor splice sites in MAdCAM-1 variants are shown. Potential 5' splice donor and 3' splice acceptor sequences identified in the four MAdCAM-1 splice variants have been aligned (SEQ ID NOS:56-59, respectively). Sequences retained are emboldened, whereas sequences deleted are in normal type. The sequences of the 3' acceptor sites conform well to the consensus for splice junctions, whereas the 5' splice donor sequences vary from the consensus for splice junctions.

FIG. 12 shows proposed structures for MAdCAM-1 splice variants. The Ig domains are shown as ovals, and the mucin domains are represented as decorated rods, where the minor mucin domain is less decorated.

FIG. 13 shows the DNA sequence of the 5'-flanking region of the human MAdCAM-1 gene (SEQ ID NO:33) and comparison with the mouse MAdCAM-1 promoter (SEQ ID NO:48). Numbers refer to nucleotide positions and are relative to the translational start codon, which is underlined. Potential transcriptional factor binding sites identified in the human and mouse 5'-flanking regions are underlined. Identical nucleotides shared by the human and mouse sequences are denoted by vertical lines.

FIGS. 14A, 14B and 14C show that the 5'-flanking region of the human MAdCAM-1 gene has promoter activity in the human dermal endothelial cell line HMEC. Figure 14A is a schematic representation of the basic luciferase vector pGL-2/B, and the expression vectors pGL-2/B-718+ and pGL-2/B-718- derived from it, which contain a 700 bp 5'-flanking region (-718 to +20 relative to the translational start) in sense and antisense orientations, respectively. Figure 14B and 14C show the relative luciferase activity directed by the expression vectors in the human dermal endothelial cell line HMEC. The results are from two separate experiments where promoter activity is expressed as the relative photon count above the background control of cells transfected with no DNA. In 14B and 14C, cells were cultured in the presence or absence of PMA. Values are the average of duplicate experiments. RT-PCR of MAdCAM-1 and glyceraldehyde 3-phosphate dehydogenase from HMEC cells was performed with the U707 and L1072 primers generating the expected band of 386 bp.

Detailed Description of the Preferred Embodiments

The present invention provides isolated nucleic acid molecules comprising a polynucleotide encoding any one of the MAdCAM-1(a-e) polypeptides having the amino acid sequences shown in FIGS. 1-5 (SEQ ID NOs:2, 4, 6, 8, 10), respectively, which was determined by sequencing a cloned cDNA. The MAdCAM-1(a-e) proteins of the present invention share sequence homology with mouse MAdCAM-1 (FIG. 7A and 7B) (SEQ ID NO:46). The nucleotide sequence shown in FIG. 1 (SEQ ID NO:1) was obtained by sequencing the HEBBC23 clone. The nucleotide sequence shown in FIG. 2 (SEQ ID NO:3) was obtained by sequencing the HSKCW36 clone, which encodes MAdCAM-1(b), a splicing variant of the deposited cDNA clone described below. The nucleotide sequence shown in FIG. 3(SEQ ID NO:5) was obtained by sequencing the

MAdCAM-1c clone, which encodes MAdCAM-1(c), a splicing variant of the deposited cDNA clone described below. The nucleotide sequence shown in FIG. 4 (SEQ ID NO:7) was obtained by sequencing the MAdCAM-1d clone, which encodes MAdCAM-1(d), a splicing variant of the deposited cDNA clone described below. The nucleotide sequence shown in FIG. 5(SEQ ID NO:9) was obtained by sequencing the MAdCAM-1e clone, which encodes MAdCAM-1(e), a splicing variant of the deposited cDNA clone described below.

The invention also relates to isolated genomic DNA molecules comprising the 5 exons (all of which are shown in Fig. 6) which comprise the coding region of any of the MAdCAM-1 splice variants (MAdCAM-1(a-e)), as well as sequence located 5' to the start codon of the first exon, which includes the promoter for the MAdCAM-1 splice variants. A genomic clone comprising this genomic DNA was deposited on October 10, 1996, at the American Type Culture Collection, 12301 Park Lawn Drive, Rockville, Maryland 20852, and given accession number 97758. The sequence of the 5' flanking region, which includes the promoter for the genes encoding any of MAdCAM-1(a-e), is given in SEQ ID NO:33. The sequences of exons 1-5 are given in SEQ ID NOS:34-38, respectively. Example 6 gives further description of how the 5 exons shown in FIG. 6, or portions thereof, can be combined in order to generate the splice variants of MAd-CAM-1.

The present invention also relates to isolated nucleic acid molecules comprising a polynucleotide encoding the MAdCAM-1(a) polypeptide encoded by the cDNA clone deposited in a bacterial host as ATCC Deposit Number 97759 on October 10, 1996. The deposited clone is contained in the pBluescript SK(-) plasmid (Stratagene, LaJolla, CA).

Nucleic Acid Molecules

Unless otherwise indicated, all nucleotide sequences determined by sequencing a DNA molecule herein were determined using an automated DNA sequencer (such as the Model 373 from Applied Biosystems, Inc.), and all amino acid sequences of polypeptides encoded by DNA molecules determined herein were predicted by translation of a DNA sequence determined as above. Therefore, as is known in the art for any DNA sequence determined by this automated approach, any nucleotide sequence determined herein may contain some errors. Nucleotide sequences determined by automation are typically at least about 90% identical, more typically at least about 95% to at least about 99.9% identical to the actual nucleotide sequence of the sequenced DNA molecule. The actual sequence can be more precisely determined by other approaches including manual DNA sequencing methods well known in the art. As is also known in the art, a single insertion or deletion in a determined nucleotide sequence compared to the actual sequence will cause a frame shift in translation of the nucleotide sequence such that the predicted amino acid sequence encoded by a determined nucleotide sequence will be completely different from the amino acid sequence actually encoded by the sequenced DNA molecule, beginning at the point of such an insertion or deletion.

Using the information provided herein, such as the nucleotide sequence in FIGS. 1-6, a nucleic acid molecule of the present invention encoding any of the MAdCAM-1(a-e) polypeptides may be obtained using standard cloning and screening procedures, such as those for cloning cDNAs using mRNA as starting material. Illustrative of the invention, the nucleic acid molecules described in FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9) were discovered in a cDNA library derived from human fetal brain cells. The genes were also identified in cDNA libraries from the following tissues: small intestine, colon, spleen, and pancreas. The determined nucleotide sequences of the MAdCAM-1(a-e) cDNAs of FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9), respectively, contain an open reading frame encoding

a protein of 382, 366, 263, 310, and 289 amino acid residues, respectively, wherein each of MAdCAM-1(a-e) has an initiation codon at positions 1-3 of their respective nucleotide sequence in FIGS. 1-5 (SEQ ID NOs: 1, 3, 5, 7, 9), and each has a predicted leader sequence of about 17 amino acid residues. The mature MAdCAM-1(a-e) polypeptides will of course lack this leader sequence. The deduced molecular weights of complete MAdCAM-1(a-e) polypeptides are about 40, 38, 27, 32 and 32.4 kDa, respectively.

In another aspect, the invention relates to isolated genomic DNA molecules comprising the 5 exons which comprise the coding region of any of the MAdCAM-1 splice variants (MAdCAM-1(a-e)), as well as sequence located 5' to the start codon of the first exon, which includes the promoter for the MAdCAM-1 splice variants. The sequence of the 5' flanking region, which includes the promoter for the genes encoding any of MAdCAM-1(a-e), is given in SEQ ID NO:33. The sequences of exons 1-5 are given in SEQ ID NOS:34-38, respectively.

In another aspect, the invention provides isolated nucleic acid molecules comprising the genomic DNA sequence contained in the clone deposited as ATCC Deposit No. 97758 on October 10, 1996.

The present invention also relates to isolated nucleic acid molecules comprising a polynucleotide encoding the MAdCAM-1(a) polypeptide encoded by the cDNA clone deposited in a bacterial host as ATCC Deposit Number 97759 on October 10, 1996. The nucleotide sequence determined by sequencing the deposited cDNA clone, MAdCAM-1(a), which is shown in FIG. 1 (SEQ ID NO:1), contains an open reading frame encoding a polypeptide of 382 amino acid residues, including an initiation codon at nucleotide positions 1-3, with a leader sequence of about 17 amino acid residues, and a predicted molecular weight of about 40 kDa. The amino acid sequence of the mature MAdCAM-1(a) protein is shown in FIG. 1, amino acid residues 18-382 (SEQ ID NO:2).

As indicated, the present invention also provides the mature form(s) of the MAdCAM-1(a-e) proteins of the present invention. According to the signal

hypothesis, proteins secreted by mammalian cells have a signal or secretory leader sequence which is cleaved from the mature protein once export of the growing protein chain across the rough endoplasmic reticulum has been initiated. Most mammalian cells and even insect cells cleave secreted proteins with the same specificity. However, in some cases, cleavage of a secreted protein is not entirely uniform, which results in two or more mature species on the protein. Further, it has long been known that the cleavage specificity of a secreted protein is ultimately determined by the primary structure of the complete protein, that is, it is inherent in the amino acid sequence of the polypeptide. Therefore, the present invention provides a nucleotide sequence encoding the mature MAdCAM-1(a-e) polypeptides having the amino acid sequence encoded by the cDNA clone shown in Figures 1-5 (SEQ ID NO:2, 4, 6, 8, 10). By the mature MAdCAM-1(a-e) proteins shown in FIGS. 1-5 is meant the mature form(s) of the MAdCAM-1 proteins produced by expression in a mammalian cell (e.g., COS cells, as described below) of the complete open reading frame encoded by the human DNA sequence of the cDNA clone contained in the vector in the deposited host. As indicated below, the actual mature MAdCAM-1(a-e) polypeptides may or may not differ from the predicted "mature" MAdCAM-1(a-e) polypeptides shown in FIGS 1-5, depending on the accuracy of the predicted cleavage site based on computer analysis.

Methods for predicting whether a protein has a secretory leader as well as the cleavage point for that leader sequence are available. For instance, the methods of McGeoch (*Virus Res. 3*:271-286 (1985)) and von Heinje (*Nucleic Acids Res. 14*:4683-4690 (1986)) can be used. The accuracy of predicting the cleavage points of known mammalian secretory proteins for each of these methods is in the range of 75-80%. von Heinje, *supra*. However, the two methods do not always produce the same predicted cleavage point(s) for a given protein.

In the present case, the predicted amino acid sequence of the complete MAdCAM-1 (a-e) polypeptides of the present invention were analyzed by a

computer program ("PSORT") (K. Nakai and M. Kanehisa, *Genomics 14*:897-911 (1992)), which is an expert system for predicting the cellular location of a protein based on the amino acid sequence. As part of this computational prediction of localization, the methods of McGeoch and von Heinje are incorporated. The analysis by the PSORT program predicted the cleavage sites between amino acids 1 7 and 18 in Figures 1-5 (SEQ ID NOS:2, 4, 6, 8, 10). Thereafter, the complete amino acid sequences were further analyzed by visual inspection, applying a simple form of the (-1,-3) rule of von Heine. von Heinje, *supra.* Thus, the leader sequence for any of the nativeMAdCAM-1(a-e) proteins is predicted to consist of amino acid residues 1-17 in Figures 1-5 (SEQ ID NOS:2, 4, 6, 8, 10), while the predicted mature native MAdCAM-1(a-e) proteins begin at residue 18.

As one of ordinary skill would appreciate, due to sequencing errors, the predicted leader sequence of the MAdCAM-1(a-e) proteins of the present invention are predicted to be about 17 amino acids in length, but may be anywhere in the range of about 14 to about 22 amino acids.

As one of ordinary skill would appreciate, due to the possibilities of sequencing errors discussed above, as well as the variability of cleavage sites for leaders in different known proteins, the predicted polypeptide corresponding to MAdCAM-1(a) comprises about 382 amino acids, but may be anywhere in the range of 368-396 amino acids. The predicted polypeptide corresponding to MAdCAM-1(b) comprises about 366 amino acids, but may be anywhere in the range of 348-382 amino acids. The predicted polypeptide corresponding to MAdCAM-1(c) comprises about 263 amino acids, but may be anywhere in the range of 250-276 amino acids. The predicted polypeptide corresponding to MAdCAM-1(d) comprises about 310 amino acids, but may be anywhere in the range of 294-325 amino acids. The predicted polypeptide corresponding to MAdCAM-1(e) comprises about 289 amino acids, but may be anywhere in the range of 275-304 amino acids.

As indicated, nucleic acid molecules of the present invention may be in the form of RNA, such as mRNA, or in the form of DNA, including, for instance,

cDNA and genomic DNA obtained by cloning or produced synthetically. The DNA may be double-stranded or single-stranded. Single-stranded DNA or RNA may be the coding strand, also known as the sense strand, or it may be the non-coding strand, also referred to as the anti-sense strand.

By "isolated" nucleic acid molecule(s) is intended a nucleic acid molecule, DNA or RNA, which has been removed from its native environment For example, recombinant DNA molecules contained in a vector are considered isolated for the purposes of the present invention. Further examples of isolated DNA molecules include recombinant DNA molecules maintained in heterologous host cells or purified (partially or substantially) DNA molecules in solution. Isolated RNA molecules include *in vivo* or *in vitro* RNA transcripts of the DNA molecules of the present invention. Isolated nucleic acid molecules according to the present invention further include such molecules produced synthetically.

Isolated nucleic acid molecules of the present invention include DNA molecules comprising an open reading frame (ORF) shown in FIGS. 1-5 (SEQ ID NOs: 1, 3, 5, 7, 9), respectively; a DNA molecule comprising the coding sequence for the mature MAdCAM-1(a) protein shown in FIG. 1 (last 365 amino acids) (SEQ ID NO:2); a DNA molecule comprising the coding sequence for the mature MAdCAM-1(b) protein shown in FIG. 2 (last 349 amino acids) (SEQ ID NO:4); a DNA molecule comprising the coding sequence for the mature MAdCAM-1(c) protein shown in FIG. 3 (last 246 amino acids) (SEQ ID NO:6); a DNA molecule comprising the coding sequence for the mature MAdCAM-1(d) protein shown in FIG. 4 (last 293 amino acids) (SEQ ID NO:8); and a DNA molecule comprising the coding sequence for the mature MAdCAM-1(e) protein shown in FIG. 5 (last 272 amino acids) (SEQ ID NO:10). The invention also includes DNA molecules which comprise a sequence substantially different from those described above but which, due to the degeneracy of the genetic code, still encode any of the MAdCAM-1(a-e) proteins. Of course, the genetic code is well known in the art. Thus, it would be routine for one skilled in the art to generate such degenerate variants.

The invention further provides an isolated nucleic acid molecule having the nucleotide sequence shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively), or a nucleic acid molecule having a sequence complementary to one of the above sequences. Such isolated molecules, particularly DNA molecules, are useful as probes for gene mapping, by *in situ* hybridization with chromosomes, and for detecting expression of the MAdCAM-1(a-e) genes in human tissue, for instance, by northern blot analysis.

The present invention is further directed to fragments of the isolated nucleic acid molecules described herein. By a fragment of an isolated nucleic acid molecule having the nucleotide sequence of the nucleotide sequences shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively), is intended fragments at least about 15 nt, and more preferably at least about 20 nt, still more preferably at least about 30 nt, and even more preferably, at least about 40 nt in length which are useful as diagnostic probes and primers as discussed herein. Of course, larger fragments 50-1150 nt in length are also useful according to the present invention as are fragments corresponding to most, if not all, of the nucleotide sequence shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively). By a fragment at least 20 nt in length, for example, is intended fragments which include 20 or more contiguous bases from the nucleotide sequence of the nucleotide sequences as shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively).

Preferred nucleic acid fragments of the present invention include nucleic acid molecules encoding epitope-bearing portions, or the transmembrane domain, or the extracellular domain, or the intracellular domain, of the MAdCAM-1(a-e) proteins. In particular, such nucleic acid fragments of the present invention include nucleic acid molecules encoding: a polypeptide comprising amino acid residues from about 52 to about 80 in FIG. 1 (SEQ ID NO:2); a polypeptide comprising amino acid residues from about 164 to about 196 in FIG. 1 (SEQ ID NO:2); and a polypeptide comprising amino acid residues from about 278 to about 321 in FIG. 1 (SEQ ID NO:2). (The inventors have determined that the

above polypeptide fragments are antigenic regions of the MAdCAM-1(a-e) proteins. Methods for determining other such epitope-bearing portions of the MAdCAM-1(a-e) proteins are described in detail below). Other preferred nucleic acid fragments include the genomic region 5' to the MAdCAM-1 gene (nucleotides residue 1 through 718 of SEQ ID NO:33), and fragments which correspond to exon 1 (nucleotide residues 1-52 of SEQ ID NO:34), exon 2 (nucleotide residues 11-295 of SEQ ID NO:35), exon 3 (nucleotide residues 11-340 of SEQ ID NO:36), exon 4 (nucleotide residues 11-343 of SEQ ID NO:37), and exon 5 (nucleotide residues 11-608 of SEQ ID NO:38) all of which are shown in FIG. 6. Knowledge of the exon-intron boundaries (see FIG 6 and Example 6). which clearly mark functional domains in the molecule, will be helpful in designing variant forms of MAdCAM-1 for use in therapy (see below).

In another aspect, the invention provides an isolated nucleic acid molecule comprising a polynucleotide which hybridizes under stringent hybridization conditions to a portion of the polynucleotide in a nucleic acid molecule of the invention described above. By "stringent hybridization conditions" is intended overnight incubation at 42°C in a solution comprising: 50% formamide, 5x SSC (150 mM NaCl, 15mM trisodium citrate), 50 mM sodium phosphate (pH 7.6), 5x Denhardt's solution, 10% dextran sulfate, and 20 g/ml denatured, sheared salmon sperm DNA, followed by washing the filters in 0.1x SSC at about 65°C.

By a polynucleotide which hybridizes to a "portion" of a polynucleotide is intended a polynucleotide (either DNA or RNA) hybridizing to at least about 15 nucleotides (nt), and more preferably at least about 20 nt, still more preferably at least about 30 nt, and even more preferably about 30-70 nt of the reference polynucleotide. These are useful as diagnostic probes and primers as discussed above and in more detail below.

By a portion of a polynucleotide of "at least 20 nt in length," for example, is intended 20 or more contiguous nucleotides from the nucleotide sequence of the reference polynucleotide (e.g., the nucleotide sequences as shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively)).

Of course, a polynucleotide which hybridizes only to a poly A sequence (such as the 3′ terminal poly(A) tract of any of the MAdCAM-1(a-e) cDNAs shown in FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9, respectively)), or to a complementary stretch of T (or U) resides, would not be included in a polynucleotide of the invention used to hybridize to a portion of a nucleic acid of the invention, since such a polynucleotide would hybridize to any nucleic acid molecule containing a poly (A) stretch or the complement thereof (e.g., practically any double-stranded cDNA clone).

As indicated, nucleic acid molecules of the present invention which encode any of the MAdCAM-1(a-e) polypeptides may include, but are not limited to, those encoding the amino acid sequence of the mature polypeptides, by themselves; the coding sequence for the mature polypeptides and additional sequences, such as those encoding the about 17 amino acid leader or secretory sequence, such as a pre-, or pro- or prepro-protein sequence; the coding sequence of the mature polypeptide, with or without the aforementioned additional coding sequences, together with additional, non-coding sequences, including for example, but not limited to introns and non-coding 5′ and 3′ sequences, such as the transcribed, non-translated sequences that play a role in transcription, mRNA processing, including splicing and polyadenylation signals, for example - ribosome binding and stability of mRNA; an additional coding sequence which codes for additional amino acids, such as those which provide additional functionalities. Thus, the sequence encoding the polypeptide may be fused to a marker sequence, such as a sequence encoding a peptide which facilitates purification of the fused polypeptide. In certain preferred embodiments of this aspect of the invention, the marker amino acid sequence is a hexa-histidine peptide, such as the tag provided in a pQE vector (Qiagen, Inc.), among others, many of which are commercially available. As described in Gentz *et al., Proc. Natl. Acad. Sci. USA 86*:821-824 (1989), for instance, hexa-histidine provides for convenient purification of the fusion protein. The "HA" tag is another peptide useful for purification which corresponds to an epitope derived from the influenza

hemagglutinin protein, which has been described by Wilson *et al., Cell 37*: 767 (1984). As discussed below, other such fusion proteins include any of the MAdCAM-1(a-e) polypeptides fused to Fc at the N- or C-terminus.

The present invention further relates to variants of the nucleic acid molecules of the present invention, which encode portions, analogs or derivatives of the MAdCAM-1(a-e) proteins. Variants may occur naturally, such as a natural allelic variant. By an "allelic variant" is intended one of several alternate forms of a gene occupying a given locus on a chromosome of an organism. *Genes II*, Lewin, B., ed., John Wiley & Sons, New York (1985). Non-naturally occurring variants may be produced using art-known mutagenesis techniques.

Such variants include those produced by nucleotide substitutions, deletions or additions, which may involve one or more nucleotides. The variants may be altered in coding regions, non-coding regions, or both. Alterations in the coding regions may produce conservative or non-conservative amino acid substitutions, deletions or additions. Especially preferred among these are silent substitutions, additions and deletions, which do not alter the properties and activities of the MAdCAM-1(a-e) proteins or portions thereof. Also especially preferred in this regard are conservative substitutions.

Further embodiments of the invention include isolated nucleic acid molecules comprising a polynucleotide having a nucleotide sequence at least 90% identical, and more preferably at least 95%, 96%, 97%, 98% or 99% identical to (a) a nucleotide sequences encoding the full-length MAdCAM-1(a-e) polypeptides having the complete amino acid sequence in FIGS. 1-5 (SEQ ID NOs:2, 4, 6, 8, 10, respectively), including the predicted leader sequence; (b) a nucleotide sequence encoding the mature MAdCAM-1(a-e) polypeptides (full-length polypeptide with the leader removed) having the amino acid sequences at positions 18-382 in FIG. 1 (SEQ ID NO:2), 18-366 in FIG. 2 (SEQ ID NO:4), 18-263 in FIG. 3 (SEQ ID NO:6), 18-310 in FIG. 4 (SEQ ID NO:8), or 18-290 in FIG. 5 (SEQ ID NO:10); (c) a nucleotide sequence encoding a polypeptide comprising the transmembrane domain of any of the MAdCAM-1

polypeptides (MAdCAM-1(a-e)); (d) a nucleotide sequence encoding a polypeptide comprising the extracellular domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e)); (e) a nucleotide sequence encoding a polypeptide comprising the intracellular domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e)); (f) a nucleotide sequence comprising the MAdCAM-1 promoter, wherein the nucleotide sequence is given in SEQ ID NO:33; (g) a nucleotide sequence encoding exon 1, 2, 3, 4 or 5 of MAdCAM-1, having the sequence given in SEQ ID NOS:34, 35, 36, 37 and 38, respectively; and (h) a nucleotide sequence complementary to any of the nucleotide sequences in (a), (b), (c), (d), (e), (f) or (g), above.

By a polynucleotide having a nucleotide sequence at least, for example, 95% "identical" to a reference nucleotide sequence encoding any of the MAdCAM-1(a-e) polypeptides is intended that the nucleotide sequence of the polynucleotide is identical to the reference sequence except that the polynucleotide sequence may include up to five point mutations per each 100 nucleotides of the reference nucleotide sequence encoding any of the MAdCAM-1(a-e) polypeptides. In other words, to obtain a polynucleotide having a nucleotide sequence at least 95% identical to a reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence may be inserted into the reference sequence. These mutations of the reference sequence may occur at the 5′ or 3′ terminal positions of the reference nucleotide sequence or anywhere between those terminal positions, interspersed either individually among nucleotides in the reference sequence or in one or more contiguous groups within the reference sequence.

As a practical matter, whether any particular nucleic acid molecule is at least 90%, 95%, 96%, 97%, 98% or 99% identical to, for instance, the nucleotide sequences shown in FIGS. 1-6 or to the nucleotides sequence of the deposited genomic clone, or to the deposited cDNA clone, can be determined

conventionally using known computer programs such as the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711). Bestfit uses the local homology algorithm of Smith and Waterman, *Advances in Applied Mathematics 2*: 482-489 (1981), to find the best segment of homology between two sequences. When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present invention, the parameters are set, of course, such that the percentage of identity is calculated over the full length of the reference nucleotide sequence and that gaps in homology of up to 5% of the total number of nucleotides in the reference sequence are allowed.

The present application is directed to nucleic acid molecules at least 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleic acid sequences shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively), or to the nucleic acid sequence of the deposited genomic DNA, irrespective of whether they encode a polypeptide having the activity of any of MAdCAM-1(a-e). This is because even where a particular nucleic acid molecule does not encode a polypeptide having MAdCAM-1(a-e) activity, one of skill in the art would still know how to use the nucleic acid molecule, for instance, as a hybridization probe or a polymerase chain reaction (PCR) primer. Uses of the nucleic acid molecules of the present invention that do not encode a polypeptide having the activity of any of MAdCAM-1(a-e) include, *inter alia,* (1) isolating the gene encoding MAdCAM-1(a-e) or allelic variants thereof in a cDNA library; (2) *in situ* hybridization (e.g., "FISH") to metaphase chromosomal spreads to provide precise chromosomal location of the gene encoding MAdCAM-1(a-e), as described in Verma *et al., Human Chromosomes: A Manual of Basic Techniques*, Pergamon Press, New York (1988); and Northern Blot analysis for detecting mRNA expression of any of MAdCAM-1(a-e) in specific tissues.

Preferred, however, are nucleic acid molecules having sequences at least 90%, 95%, 96%, 97%, 98% or 99% identical to any of the nucleic acid sequences shown in FIGS. 1-6 (SEQ ID NOs:1, 3, 5, 7, 9, 33, 34, 35, 36, 37, and 38, respectively), or to the nucleic acid sequence of the deposited genomic DNA which does, in fact, encode a polypeptide having the protein activity of any of MAdCAM-1(a-e). By "a polypeptide having the protein activity of any of MAdCAM-1(a-e)" is intended polypeptides exhibiting activity similar, but not necessarily identical, to an activity of any of the MAdCAM-1(a-e) proteins of the invention (either the full-length proteins or, preferably, the mature proteins), as measured in a particular biological assay. For example, the protein activity of any of MAdCAM-1(a-e) can be measured by using a variation of the Stamper-Woodruff *in vitro* lymphocyte-endothelial cell binding assay (*J. Exp. Med. 144*: 828-833 (1976), which tests the ability of lymphoid cells expressing the $\alpha_4\beta_7$ to bind to vascular endothelial cells expressing a polypeptide suspected of having the activity of any of the MAdCAM-1(a-e) proteins (Hanninen *et al., J. Clin. Invest. 92*: 2590-2515 (1993). Briefly, the assay involves contacting a cell which expresses $\alpha_4\beta_7$ (such as TK1 cells) and thus binds to cells expressing any of MAdCAM-1(a-e), with cells expressing any of the MAdCAM-1(a-e) molecules of the invention, and measuring the resultant adhesion between the two types of cells. Thus, a cell expressing the protein activity of any of MAdCAM-1(a-e) will bind to the cells expressing $\alpha_4\beta_7$, while a cell expressing a protein which does not bind to $\alpha_4\beta_7$ will be considered not to have the activity of any of MAdCAM-1(a-e).

Of course, due to the degeneracy of the genetic code, one of ordinary skill in the art will immediately recognize that a large number of the nucleic acid molecules having a sequence at least 90%, 95%, 96%, 97%, 98%, or 99% identical to the nucleic acid sequences shown in FIGS. 1-5 (SEQ ID NO:1, 3, 5, 7, 9, respectively) will encode a polypeptide "having the protein activity of any of MAdCAM-1(a-e)." In fact, since degenerate variants of these nucleotide sequences all encode the same polypeptide, this will be clear to the skilled artisan

even without performing the above described comparison assay. It will be further recognized in the art that, for such nucleic acid molecules that are not degenerate variants, a reasonable number will also encode a polypeptide having the protein activity of any of MAdCAM-1(a-e). This is because the skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly effect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid).

For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided in Bowie, J. U. *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247:*1306-1310 (1990), wherein the authors indicate that proteins are surprisingly tolerant of amino acid substitutions.

Vectors and Host Cells

The present invention also relates to vectors which include the isolated DNA molecules of the present invention, host cells which are genetically engineered with the recombinant vectors, and the production of any of the MAdCAM-1(a-e) polypeptides or fragments thereof by recombinant techniques.

The polynucleotides may be joined to a vector containing a selectable marker for propagation in a host. Generally, a plasmid vector is introduced in a precipitate, such as a calcium phosphate precipitate, or in a complex with a charged lipid. If the vector is a virus, it may be packaged *in vitro* using an appropriate packaging cell line and then transduced into host cells.

The DNA insert should be operatively linked to an appropriate promoter, such as the phage lambda PL promoter, the *E. coli lac, trp* and *tac* promoters, the SV40 early and late promoters and promoters of retroviral LTRs, to name a few. Other suitable promoters will be known to the skilled artisan. The expression constructs will further contain sites for transcription initiation, termination and, in the transcribed region, a ribosome binding site for translation. The coding

portion of the mature transcripts expressed by the constructs will preferably include a translation initiating at the beginning and a termination codon (UAA, UGA or UAG) appropriately positioned at the end of the polypeptide to be translated.

As indicated, the expression vectors will preferably include at least one selectable marker. Such markers include dihydrofolate reductase or neomycin resistance for eukaryotic cell culture and tetracycline or ampicillin resistance genes for culturing in *E. coli* and other bacteria. Representative examples of appropriate hosts include, but are not limited to, bacterial cells, such as *E. coli*, *Streptomyces* and *Salmonella typhimurium* cells; fungal cells, such as yeast cells; insect cells such as *Drosophila* S2 and *Spodoptera* Sf9 cells; animal cells such as CHO, COS and Bowes melanoma cells; and plant cells. Appropriate culture mediums and conditions for the above-described host cells are known in the art.

Among vectors preferred for use in bacteria include pQE70, pQE60 and pQE-9, available from Qiagen; pBS vectors, Phagescript vectors, Bluescript vectors, pNH8A, pNH16a, pNH18A, pNH46A, available from Stratagene; and ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 available from Pharmacia. Among preferred eukaryotic vectors are pWLNEO, pSV2CAT, pOG44, pXT1 and pSG available from Stratagene; and pSVK3, pBPV, pMSG and pSVL available from Pharmacia. Other suitable vectors will be readily apparent to the skilled artisan.

Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-dextran mediated transfection, cationic lipid-mediated transfection, electroporation, transduction, infection or other methods. Such methods are described in many standard laboratory manuals, such as Davis *et al.*, *Basic Methods In Molecular Biology* (1986).

The polypeptide may be expressed in a modified form, such as a fusion protein, and may include not only secretion signals, but also additional heterologous functional regions. For instance, a region of additional amino acids, particularly charged amino acids, may be added to the N-terminus of the

polypeptide to improve stability and persistence in the host cell, during purification, or during subsequent handling and storage. Also, peptide moieties may be added to the polypeptide to facilitate purification. Such regions may be removed prior to final preparation of the polypeptide. The addition of peptide moieties to polypeptides to engender secretion or excretion, to improve stability and to facilitate purification, among others, are familiar and routine techniques in the art. A preferred fusion protein comprises a heterologous region from immunoglobulin that is useful to solubilize proteins. For example, EP-A-O 464 533 (Canadian counterpart 2045869) discloses fusion proteins comprising various portions of constant region of immunoglobulin molecules together with another human protein or part thereof. In many cases, the Fc part in a fusion protein is thoroughly advantageous for use in therapy and diagnosis and thus results, for example, in improved pharmacokinetic properties (EP-A 0232 262). On the other hand, for some uses it would be desirable to be able to delete the Fc part after the fusion protein has been expressed, detected and purified in the advantageous manner described. This is the case when Fc portion proves to be a hindrance to use in therapy and diagnosis, for example when the fusion protein is to be used as antigen for immunizations. In drug discovery, for example, human proteins, such as, hIL5- has been fused with Fc portions for the purpose of high-throughput screening assays to identify antagonists of hIL-5. See, D. Bennett et al., *Journal of Molecular Recognition*, Vol. 8 52-58 (1995) and K. Johanson et al., *The Journal of Biological Chemistry*, Vol. 270, No. 16, pp 9459-9471 (1995).

The MAdCAM-1(a-e) proteins can be recovered and purified from recombinant cell cultures by well-known methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Most preferably, high performance liquid chromatography ("HPLC") is employed for purification. Polypeptides of the present invention include naturally purified products, products of chemical

synthetic procedures, and products produced by recombinant techniques from a prokaryotic or eukaryotic host, including, for example, bacterial, yeast, higher plant, insect and mammalian cells. Depending upon the host employed in a recombinant production procedure, the polypeptides of the present invention may be glycosylated or may be non-glycosylated. In addition, polypeptides of the invention may also include an initial modified methionine residue, in some cases as a result of host-mediated processes.

MAdCAM-1(a-e) Polypeptides and Fragments

The invention further provides isolated MAdCAM-1(a-e) polypeptides having the amino acid sequence given in FIG. 1-5 (SEQ ID NO:2, 4, 6, 8, 10, respectively), or a peptide or polypeptide comprising a portion of the above polypeptides, as well as any of the polypeptides encoded by the nucleotide sequence of exons 1-5 of FIG 6 (SEQ ID NOS:34-38).

It will be recognized in the art that some amino acid sequences of the MAdCAM-1(a-e) polypeptides can be varied without significant effect of the structure or function of the protein. If such differences in sequence are contemplated, it should be remembered that there will be critical areas on the protein which determine activity.

Thus, the invention further includes variations of the MAdCAM-1(a-e) polypeptides which show substantial MAdCAM-1(a-e) polypeptide activity or which include regions of any of the MAdCAM-1(a-e) proteins such as the protein portions discussed below. Such mutants include deletions, insertions, inversions, repeats, and type substitutions. As indicated above, further guidance concerning which amino acid changes are likely to be phenotypically silent can be found in Bowie, J.U., *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247:*1306-1310 (1990).

Thus, the fragment, derivative or analog of the polypeptide shown in FIGS 1-5 (SEQ ID NOS: 2, 4, 6 ,8 ,10) may be (i) one in which one or more of

the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or (iv) one in which the additional amino acids are fused to the mature polypeptide, such as an IgG Fc fusion region peptide or leader or secretory sequence or a sequence which is employed for purification of the mature polypeptide or a proprotein sequence. Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

Of particular interest are substitutions of charged amino acids with another charged amino acid and with neutral or negatively charged amino acids. The latter results in proteins with reduced positive charge to improve the characteristics of the MAdCAM-1(a-e) proteins. The prevention of aggregation is highly desirable. Aggregation of proteins not only results in a loss of activity but can also be problematic when preparing pharmaceutical formulations, because they can be immunogenic. (Pinckard *et al.*, *Clin Exp. Immunol.* 2:331-340 (1967); Robbins *et al.*, *Diabetes* 36:838-845 (1987); Cleland *et al. Crit. Rev. Therapeutic Drug Carrier Systems* 10:307-377 (1993)).

As indicated, changes are preferably of a minor nature, such as conservative amino acid substitutions that do not significantly affect the folding or activity of the protein (see Table 1).

TABLE 1. Conservative Amino Acid Substitutions.

Aromatic	Phenylalanine Tryptophan Tyrosine
Hydrophobic	Leucine Isoleucine Valine
Polar	Glutamine Asparagine
Basic	Arginine Lysine Histidine
Acidic	Aspartic Acid Glutamic Acid
Small	Alanine Serine Threonine Methionine Glycine

Amino acids in the MAdCAM-1(a-e) polypeptides of the present invention that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (Cunningham and Wells, *Science 244*:1081-1085 (1989)). The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity such as receptor binding or *in vitro*, or *in vitro* proliferative activity. Sites that are critical for protein activity can also be determined by structural analysis such as crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith *et al.*, *J. Mol. Biol. 224*:899-904 (1992) and de Vos *et al. Science 255*:306-312 (1992)).

The polypeptides of the present invention are preferably provided in an isolated form, and preferably are substantially purified. A recombinantly produced version of any of the MAdCAM-1(a-e) polypeptides can be

substantially purified by the one-step method described in Smith and Johnson, *Gene 67*:31-40 (1988).

The polypeptides of the present invention include any of the polypeptides of FIGS. 1-5(SEQ ID NOS:2, 4, 6, 8, 10, respectively) including the leader, any of the mature polypeptides of FIGS. 1-5 (SEQ ID NOS:2, 4, 6, 8, 10, respectively) minus the leader (i.e., the mature protein), any of the polypeptides of FIGS. 1-5(SEQ ID NOS:2, 4, 6, 8, 10, respectively) minus the leader, the extracellular domain of any of the polypeptides of FIGS. 1-5(SEQ ID NOS:2, 4, 6, 8, 10, respectively), the intracellular domain of any of the polypeptides of FIGS. 1-5(SEQ ID NOS:2, 4, 6, 8, 10, respectively), and the transmembrane domain of any of the polypeptides of FIGS. 1-5(SEQ ID NOS:2, 4, 6, 8, 10, respectively), as well as any of the polypeptides encoded by the nucleotide sequence of exons 1-5 of FIG 6 (SEQ ID NOS:34-38). Of course, those of ordinary skill will understand that, just as the splicing variants MAdCAM-1(a-e) are generated in vivo by alternative splicing of the 5 exons shown in FIG. 6 (SEQ ID NOS:34-38) (as well as by splicing internal to those exons, see Example 6), polypeptide variants of MAdCAM-1 can be recombinantly prepared by combining exons, or portions of exons, of the sequences shown in FIG. 6 (SEQ ID NOS:34-38). Such polypeptides are also included in the invention. Also included are polypeptides which are at least 80% identical, more preferably at least 90% or 95% identical, still more preferably at least 96%, 97%, 98% or 99% identical to the above-mentioned polypeptides, and also include portions of such polypeptides with at least 30 amino acids and more preferably at least 50 amino acids.

By a polypeptide having an amino acid sequence at least, for example, 95% "identical" to a reference amino acid sequence of any of the MAdCAM-1(a-e) polypeptides is intended that the amino acid sequence of the polypeptide is identical to the reference sequence except that the polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the reference amino acid of any of the MAdCAM-1(a-e) polypeptides. In

other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a reference amino acid sequence, up to 5% of the amino acid residues in the reference sequence may be deleted or substituted with another amino acid, or a number of amino acids up to 5% of the total amino acid residues in the reference sequence may be inserted into the reference sequence. These alterations of the reference sequence may occur at the amino or carboxy terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence or in one or more contiguous groups within the reference sequence.

As a practical matter, whether any particular polypeptide is at least 90%, 95%, 96%, 97%, 98% or 99% identical to, for instance, any of the amino acid sequences shown in FIGS. 1-6 (SEQ ID NOs:2, 4, 6, 8, 10, respectively), or to the amino acid sequence encoded by deposited genomic DNA, can be determined conventionally using known computer programs such the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711). When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present invention, the parameters are set, of course, such that the percentage of identity is calculated over the full length of the reference amino acid sequence and that gaps in homology of up to 5% of the total number of amino acid residues in the reference sequence are allowed.

The polypeptide of the present invention could be used as a molecular weight marker on SDS-PAGE gels or on molecular sieve gel filtration columns using methods well known to those of skill in the art.

In another aspect, the invention provides a peptide or polypeptide comprising an epitope-bearing portion of the invention described hererin. The epitope of this polypeptide portion is an immunogenic or antigenic epitope of a polypeptide of the invention. An "immunogenic epitope" is defined as a part of a protein that elicits an antibody response when the whole protein is the

immunogen. On the other hand, a region of a protein molecule to which an antibody can bind is defined as an "antigenic epitope." The number of immunogenic epitopes of a protein generally is less than the number of antigenic epitopes. See, for instance, Geysen *et al., Proc. Natl. Acad. Sci. USA 81*:3998-4002 (1983).

As to the selection of peptides or polypeptides bearing an antigenic epitope (i.e., that contain a region of a protein molecule to which an antibody can bind), it is well known in that art that relatively short synthetic peptides that mimic part of a protein sequence are routinely capable of eliciting an antiserum that reacts with the partially mimicked protein. See, for instance, Sutcliffe, J. G., Shinnick, T. M., Green, N. and Learner, R.A. (1983) Antibodies that react with predetermined sites on proteins. *Science 219:*660-666. Peptides capable of eliciting protein-reactive sera are frequently represented in the primary sequence of a protein, can be characterized by a set of simple chemical rules, and are confined neither to immunodominant regions of intact proteins (i.e., immunogenic epitopes) nor to the amino or carboxyl terminals.

Antigenic epitope-bearing peptides and polypeptides of the invention are therefore useful to raise antibodies, including monoclonal antibodies, that bind specifically to a polypeptide of the invention. See, for instance, Wilson *et al., Cell 37*:767-778 (1984) at 777.

Antigenic epitope-bearing peptides and polypeptides of the invention preferably contain a sequence of at least seven, more preferably at least nine and most preferably between about at least about 15 to about 30 amino acids contained within the amino acid sequence of a polypeptide of the invention.

Non-limiting examples of antigenic polypeptides or peptides that can be used to generate antibodies specific to any of the MAdCAM-1(a-e) polypeptides include: a polypeptide comprising amino acid residues from about 52 to about 80 in FIG. 1 (SEQ ID NO:2); a polypeptide comprising amino acid residues from about 164 to about 196 in FIG. 1 (SEQ ID NO:2); and a polypeptide comprising amino acid residues from about 228 to about 321 in FIG. 1 (SEQ ID NO:2). As

indicated above, the inventors have determined that the above polypeptide fragments are antigenic regions of the endokine alpha protein.

The epitope-bearing peptides and polypeptides of the invention may be produced by any conventional means. Houghten, R. A. (1985) General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. *Proc. Natl. Acad. Sci. USA 82*:5131-5135. This "Simultaneous Multiple Peptide Synthesis (SMPS)" process is further described in U.S. Patent No. 4,631,211 to Houghten et al. (1986).

MAdCAM-1 Related Disorder Diagnosis

Under circumstances which induce an inflammatory response, circulating lymphocytes expressing a receptor for one or more of the MAdCAM-1 proteins (MAdCAM-1(a-e)) are believed to bind to the MAdCAM-1 protein on mucosal venules, and then migrate through the venules to the epithelium, where acute inflammation results. Therefore, the invention also relates to the diagnosis of a pathological inflammatory condition by identifying the presence of an enhanced level of one or more of the MAdCAM-1(a-e) proteins or mRNA encoding these proteins, as compared to a corresponding "standard" mammal, i.e., a mammal of the same species not having the pathological inflammatory condition. Such conditions include transplantation rejection, arthritis, rheumatoid arthritis, infection, dermatosis, inflammatory bowel disease, and autommune disease, including chronic relapsing experimental autoimmune encephalitis (EAE).

It is also believed that certain tissues in mammals with cancer express significantly enhanced levels of one or more of the MAdCAM-1(a-e) proteins and mRNA encoding these proteins when compared to a corresponding "standard" mammal, i.e., a mammal of the same species not having the cancer. Further, it is believed that enhanced levels of any of the MAdCAM-1(a-e) proteins can be detected in certain body fluids (e.g., sera, plasma, urine, and spinal fluid) from

mammals with cancer when compared to sera from mammals of the same species not having the cancer. Thus, the invention provides a diagnostic method useful during tumor diagnosis, which involves assaying the expression level of the gene encoding any of the MAdCAM-1(a-e) proteins in mammalian cells or body fluid and comparing the gene expression level with a standard expression level for that same gene, whereby an increase in the gene expression level over the standard is indicative of certain tumors.

Where a tumor diagnosis has already been made according to conventional methods, the present invention is useful as a prognostic indicator, whereby patients exhibiting enhanced expression of any of the MAdCAM-1(a-e) genes will experience a worse clinical outcome relative to patients expressing the relevant gene at a lower level.

By "assaying the expression level of the gene encoding one or more of the MAdCAM-1(a-e) proteins" is intended qualitatively or quantitatively measuring or estimating the level of one or more of the MAdCAM-1(a-e) proteins or the level of the mRNA encoding one or more of the MAdCAM-1(a-e) proteins in a first biological sample either directly (e.g., by determining or estimating absolute protein level or mRNA level) or relatively (e.g., by comparing to the protein level or mRNA level of the same MAdCAM-1(a-e)in a second biological sample).

Preferably, the level of the MAdCAM-1(a-e) protein or mRNA level in the first biological sample is measured or estimated and compared to a standard protein level or mRNA level for the same protein, the standard being taken from a second biological sample obtained from an individual not having the cancer. As will be appreciated in the art, once a standard protein level or mRNA level for one or more of MAdCAM-1(a-e) is known, it can be used repeatedly as a standard for comparison.

By "biological sample" is intended any biological sample obtained from an individual, cell line, tissue culture, or other source which contains one or more of the MAdCAM-1(a-e)proteins or the mRNA encoding them. Biological samples include mammalian body fluids (such as sera, plasma, urine, synovial

fluid and spinal fluid) which contain a secreted mature protein, and ovarian, prostate, heart, placenta, pancreas liver, spleen, lung, breast and umbilical tissue.

The present invention is useful for detecting cancer in mammals. In particular the invention is useful during diagnosis of the of following types of cancers in mammals: lymphoma, leukemia, and metastatic tumors. Preferred mammals include monkeys, apes, cats, dogs, cows, pigs, horses, rabbits and humans. Particularly preferred are humans.

Total cellular RNA can be isolated from a biological sample using the single-step guanidinium-thiocyanate-phenol-chloroform method described in Chomczynski and Sacchi, *Anal. Biochem. 162:*156-159 (1987). Levels of mRNA encoding any of the MAdCAM-1(a-e) proteins are then assayed using any appropriate method. These include Northern blot analysis, S1 nuclease mapping, the polymerase chain reaction (PCR). reverse transcription in combination with the polymerase chain reaction (RT-PCR), and reverse transcription in combination with the ligase chain reaction (RT-LCR).

Assaying protein levels of any of MAdCAM-1(a-e) in a biological sample can occur using antibody-based techniques. For example, expression of any of the MAdCAM-1(a-e) polypeptides in tissues can be studied with classical immunohistological methods. (Jalkanen, M., *et al., J. Cell. Biol. 101:*976-985 (1985); Jalkanen, M., *et al., J. Cell . Biol. 105:*3087-3096 (1987)).

Other antibody-based methods useful for detecting MAdCAM-1(a-e) protein gene expression include immunoassays, such as the enzyme linked immunosorbent assay (ELISA) and the radioimmunoassay (RIA). Suitable labels are known in the art, and include enzyme labels, such as glucose oxidase, and radioisotopes, such as iodine (^{125}I, ^{121}I), carbon (^{14}C), sulfur (^{35}S), tritium (^{3}H), indium (^{112}In), and technetium (^{99m}Tc), and fluorescent labels, such as fluorescein and rhodamine, and biotin.

Chromosome Assays

The nucleic acid molecules of the present invention are also valuable for chromosome identification. The sequence is specifically targeted to and can hybridize with human chromosome 19p13.3. The mapping of DNAs to chromosomes according to the present invention is an important first step in correlating those sequences with genes associated with disease.

In certain preferred embodiments in this regard, the cDNA herein disclosed is used to clone genomic DNA of any of the genes encoding MAdCAM-1(a-e) proteins. This can be accomplished using a variety of well known techniques and libraries, which generally are available commercially. The genomic DNA then is used for *in situ* chromosome mapping using well known techniques for this purpose.

In addition, in some cases, sequences can be mapped to chromosomes by preparing PCR primers (preferably 15-25 bp) from the cDNA. Computer analysis of the 3′ untranslated region of the gene is used to rapidly select primers that do not span more than one exon in the genomic DNA, thus complicating the amplification process. These primers are then used for PCR screening of somatic cell hybrids containing individual human chromosomes.

Fluorescence *in situ* hybridization ("FISH") of a cDNA clone to a metaphase chromosomal spread can be used to provide a precise chromosomal location in one step. This technique can be used with probes from the cDNA as short as 50 or 60 bp. For a review of this technique, see Verma *et al., Human Chromosomes: A Manual Of Basic Techniques*, Pergamon Press, New York (1988).

Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. Such data are found, for example, in V. McKusick, *Mendelian Inheritance In Man*, available on-line through Johns Hopkins University, Welch Medical Library. The relationship between genes and diseases that have been

mapped to the same chromosomal region are then identified through linkage analysis (coinheritance of physically adjacent genes).

Next, it is necessary to determine the differences in the cDNA or genomic sequence between affected and unaffected individuals. If a mutation is observed in some or all of the affected individuals but not in any normal individuals, then the mutation is likely to be the causative agent of the disease.

MAdCAM-1 Protein and Antibody Therapy

Under circumstances which induce an inflammatory response, circulating lymphocytes are believed to express a receptor for one or more of the MAdCAM-1 proteins (MAdCAM-1(a-e)), bind to the MAdCAM-1 protein on mucosal venules via this receptor, and then migrate through the venules to the epithelium, where acute inflammation results. Therefore, the administration of a therapeutic composition capable of blocking the migration of leukocytes via MAdCAM-1 polypeptides (MAdCAM-1(a-e)) (*i.e.*, an antagonist of the activity of any of MAdCAM-1(a-e)) could be an effective therapeutic treatment for minimizing tissue damage in many abnormal inflammatory conditions, especially where the inflammation is chronic or acute. Such conditions include transplantation rejection, arthritis, rheumatoid arthritis, infection, dermatosis, inflammatory bowel disease, and autommune disease, including chronic relapsing experimental autoimmune encephalitis (EAE).

Thus, the invention also relates to a therapeutic method for treating an individual in need of a reduction in the activity of any of MAdCAM-1(a-e) by administering to the individual a therapeutically effective amount of a composition comprising an antagonist of MAdCAM-1(a-e) activity. Such compounds include anti-MAdCAM-1 antibodies or fragments thereof, as well as compounds such as solubilized $\alpha_4\beta_7$. Such individuals can include those suffering from abnormal inflammatory conditions, especially where the inflammation is chronic or acute. The invention also includes using such

compositions as a "preventative" treatment before detection of an inflammatory state, so as to prevent the development of inflammation in a patient at high risk for the same, such as, for example, transplant patients.

Therefore, the invention is further directed to antibody-based therapies which involve administering an antibody directed against any of MAdCAM-1(a-e), to a mammalian, preferably human, patient for treating one or more of the above-described disorders. Methods for producing such anti-MAdCAM-1 polyclonal and monoclonal antibodies are described in detail above. Such antibodies may be provided in pharmaceutically acceptable compositions as known in the art or as described herein.

A summary of the ways in which the antibodies of the present invention may be used therapeutically includes binding any of the MAdCAM-1(a-e) polypeptides locally or systemically in the body. Some of these approaches are described in more detail below. Armed with the teachings provided herein, one of ordinary skill in the art will know how to use the antibodies of the present invention for diagnostic, monitoring or therapeutic purposes without undue experimentation.

The antagonists of MAdCAM-1(a-e) activity of the invention may also include soluble forms of any of the MAdCAM-1(a-e) polypeptides. The administration of soluble forms of any of the MAdCAM-1(a-e) polypeptides may block leukocyte adhesion to endothelium at sites of inflammation. Those of skill in the art will readily know how to generate such soluble fragments based on an analysis of the MAdCAM-1 three dimensional structure such as that given in FIG. 7.

Modes of administration

It will be appreciated that conditions caused by an increase in the standard or normal level of activity of any of MAdCAM-1(a-e) in an individual, can be treated by administration of a molecule capable of blocking lymphocyte adhesion

that is mediated by any of MAdCAM-1(a-e). Thus, the invention further provides a method of treating an individual in need of a decreased level of MAdCAM-1(a-e)-mediated adhesion comprising administering to such an individual a pharmaceutical composition comprising an effective amount of antagonist of any of the MAdCAM-1(a-e) polypeptides of the invention. Such antagonists include anti-MAdCAM-1 antibodies or fragments or derivatives thereof, as well as compounds such as solubilized $\alpha_4\beta_7$, or soluble forms of any of MAdCAM-1(a-e), which are effective to decrease the activity level of the desired MAdCAM-1(a-e) protein in such an individual.

As a general proposition, the total pharmaceutically effective amount of one or more of the antagonists, including antibodies, soluble forms of $\alpha_4\beta_7$, and soluble forms of the MAdCAM-1(a-e) polypeptides, administered parenterally per dose will be in the range of about 1 μg/kg/day to 10 mg/kg/day of patient body weight, although, as noted above, this will be subject to therapeutic discretion. More preferably, this dose is at least 0.01 mg/kg/day, and most preferably for humans between about 0.01 and 1 mg/kg/day for the hormone. If given continuously, the desired antagonist of the MAdCAM-1(a-e) polypeptides is typically administered at a dose rate of about 1 μg/kg/hour to about 50 μg/kg/hour, either by 1-4 injections per day or by continuous subcutaneous infusions, for example, using a mini-pump. An intravenous bag solution may also be employed.

Pharmaceutical compositions containing one or more of the antagonists of the MAdCAM-1(a-e) polypeptides of the invention may be administered orally, rectally, parenterally, intracisternally, intravaginally, intraperitoneally, topically (as by powders, ointments, drops or transdermal patch), bucally, or as an oral or nasal spray. By "pharmaceutically acceptable carrier" is meant a non-toxic solid, semisolid or liquid filler, diluent, encapsulating material or formulation auxiliary of any type. The term "parenteral" as used herein refers to modes of administration which include intravenous, intramuscular,

intraperitoneal, intrasternal, subcutaneous and intraarticular injection and infusion.

Where the antagonist to be used is an antibody, fragment thereof, or derivative thereof, it is preferred to use high affinity and/or potent *in vivo* MAdCAM-1-inhibiting and/or neutralizing antibodies, fragments or regions thereof, for both MAdCAM-1 immunoassays (see the section of this application directed to diagnostics) and therapy of endokine related disorders. Such antibodies, fragments, or regions, will preferably have an affinity for any of human MAdCAM-1(a-e), expressed as Ka, of at least 10^8 M^{-1}, more preferably, at least 10^9 M^{-1}, such as 5 X 10^8 M^{-1}, 8 X 10^8 M^{-1}, 2 X 10^9 M^{-1}, 4 X 10^9 M^{-1}, 6 X 10^9 M^{-1}, 8 X 10^9 M^{-1}.

Preferred for human therapeutic use are high affinity murine and murine/human or human/human chimeric antibodies, and fragments, regions and derivatives having potent *in vivo* MAdCAM-1-inhibiting and/or neutralizing activity, according to the present invention, *e.g.*, that block MAdCAM-1-mediated cell adhesion activity, *in vivo*, *in situ*, and *in vitro*.

Selection of Compounds Capable of Regulating Expression of MAdCAM-1

As the invention also includes isolated genomic DNA molecules comprising the 5' flanking region of MAdCAM-1(a-e), including the promoter for these splice variants, yet another aspect of the invention is related to a method for identifying compounds capable of enhancing or inhibiting expression of any of MAdCAM-1(a-e). In order to determine the effect of such compounds, reporter plasmids are constructed by linking a portion of the DNA located 5' to the transcription start site of any of MAdCAM-1(a-e) in front of a reporter gene. Such constructs are then transfected into appropriate cell lines. Compounds that are to be tested for their ability to increase or decrease expression from the MAdCAM-1 promoter are then administered to the cell bearing the reporter construct, and the effect of each compound on reporter gene expression is

determined by comparing that level of expression to the expression level in a control cell bearing the reporter construct, where the test compound has not been administered to the control cell.

The DNA sequence of the 5' flanking region of the MAdCAM-1 gene is shown in Figure 6 (SEQ ID NO:33). For a full description of this region, see Example 6, below. Of course, since the nucleotide sequence is known, routine methods are available for producing such nucleic acid molecules synthetically (see, for example, *Synthesis and Application of DNA and RNA*, S.A. Narang, ed., 1987, Academic Press, San Diego, CA). Alternatively, such isolated nucleic acid molecules of the present invention can be generated as follows. The MAdCAM-1 gene promoter region is obtained by amplification using the polymerase chain reaction (PCR). The amplified fragment is then inserted into an appropriate plamid (such as, for example, pCAT ™ (Promega, Madison, WI)). Nested deletion plasmids are then generated using the commercially available "Erase-a-Base" System (Promega, Madison, WI) as described in Henikoff, *Gene 28:*351-359 (1984)). Thus, only routine experimentation would be required to generate any of the isolated nucleic acid molecules of the present invention which are capable of enhancing or inhibiting gene expression.

The nucleic acid molecules of the present invention can include the MAdCAM-1 promoter and *cis*-acting enhancer and/or silencer elements capable of affecting gene transcription. For simplicity, these isolated nucleic acid molecules of the present invention are referred to below as "MAdCAM-1 transcriptional regulatory elements" or "transcriptional elements." As indicated, to determine the effect of a transcriptional element of the present invention on gene expression, nested deletion reporter plasmids can be generated containing a transcriptional element of the present invention linked in front of the chloramphenicol acetyltransferase (CAT) reporter gene. Such recombinant DNA molecules of the present invention actually generated by the inventors include transcriptional elements inserted, in both orientations, into the

*Xba*I site of pBLCAT2 vector (Luckow, B., Schütz, G., *Nucleic Acids Res. 15*:5490 (1987)).

By the invention, a recombinant DNA molecule containing a transcriptional element of the present invention is used to transiently transfect an appropriate cell line such as, for example, human choriocarcinoma cell lines (JEG-3 and JAR), the human prostate carcinoma cell line PC-3, or the monkey kidney cell line CV-1, all of which are availabe form the American Type Culture Collection. In addition to using the CAT system for reporter gene analyses, the hGH transient expression system can also be used (Selden *et al., Mol. Cell Biol. 6*:3173-3179 (1986)) or other systems that are based on the expression of β-galactosidase (An *et al., Mol. Cell. Biol.* 2:1628-1632 (1982)) and xanthine-guanine phosphoribosyl transferase (Chu *et al., Nucleic Acids Res. 13*:2921-2930 (1985)).

A transcriptional element of the present invention may be inserted into an appropriate vector in accordance with conventional techniques, including blunt-ending or staggered-ending termini for ligation, restriction enzyme digestion to provide appropriate termini, filling in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and ligation with appropriate ligases. Techniques for such manipulations are disclosed by Maniatis, T., *et al., infra*, and are well known in the art. Clones containing a transcriptional element of the present invention may be identified by any means which specifically selects for a MAdCAM-1 enhancer or silencer region DNA such as, for example by hybridization with an appropriate nucleic acid probe(s) containing a sequence complementary to all or part of the transcriptional element. Oligonucleotide probes specific for a transcriptional element of the present invention can be designed simply by reference to SEQ ID No:33. Techniques for nucleic acid hybridization and clone identification are disclosed by Maniatis, T., *et al.*, (In: *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratories, Cold Spring Harbor, NY (1982)), and by Hames, B.D., *et al.*, (In: *Nucleic Acid Hybridization, A Practical*

Approach, IRL Press, Washington, DC (1985)). To facilitate the detection of the desired clone containing a transcriptional element of the present invention, the above-described nucleic acid probe may be labeled with a detectable group. Such detectable groups can be any material having a detectable physical or chemical property. Such materials have been well-developed in the field of nucleic acid hybridization and in general most any label useful in such methods can be applied to the present invention. Particularly useful are radioactive labels, such as ^{32}P, ^{3}H, ^{14}C, ^{35}S, ^{125}I, or the like. Any radioactive label may be employed which provides for an adequate signal and has a sufficient half-life. The oligonucleotide may be radioactively labeled, for example, by "nick-translation" by well-known means, as described in, for example, Rigby, P.J.W., *et al.*, *J. Mol. Biol. 113*:237 (1977) and by T4 DNA polymerase replacement synthesis as described in, for example, Deen, K.C., *et al.*, *Anal. Biochem. 135*:456 (1983). Alternatively, polynucleotides are also useful as nucleic acid hybridization probes when labeled with a non-radioactive marker such as biotin, an enzyme or a fluorescent group. See, for example, Leary, J.J., *et al.*, *Proc. Natl. Acad. Sci. USA 80*:4045 (1983); Renz, M., *et al.*, *Nucl. Acids Res. 12*:3435 (1984); and Renz, M., *EMBO J. 6*:817 (1983).

As used herein, "heterologous protein" is intended to refer to a peptide sequence that is heterologous to the transcriptional regulatory elements of the invention. A skilled artisan will recognize that, if desired, the teaching herein will also apply to the expression of genetic sequences encoding the MAdCAM-1 protein, or splice variants thereof, by such transcriptional regulatory elements. The reporter genes for use in the screening assay described below can code for either the MAdCAM-1 protein, or splice variants thereof, or a heterologous protein. Alternatively, detection of reporter gene expression can be at the mRNA level, such as, for example, detection of MAdCAM-1 mRNA.

To express a reporter gene under the control of the transcriptional regulatory elements of the invention, the gene must be "operably-linked" to the regulatory element. An operable linkage is a linkage in which a desired

sequence is connected to a transcriptional or translational regulatory sequence (or sequences) in such a way as to place expression (or operation) of the desired sequence under the influence or control of the regulatory sequence.

Two DNA sequences (such as a reporter gene and a promoter region sequence linked to the 5′ end of the reporter gene) are said to be operably linked if induction of promoter function results in the transcription of the reporter gene and if the nature of the linkage between the two DNA sequences does not (1) result in the introduction of a frame-shift mutation (if reporter protein activity is necessary for detection of reporter gene expression), (2) interfere with the ability of the expression regulatory sequences to direct reporter gene expression, or (3) interfere with the ability of reporter gene to be transcribed by the promoter region sequence. Thus, a promoter would be operably linked to a DNA sequence if the promoter were capable of affecting transcription of that DNA sequence.

In a similar manner, a transcriptional regulatory element of the present invention that enhances or represses gene expression may be operably-linked to such a promoter. Exact placement of the element in the nucleotide chain is not critical as long as the element is located at a position from which the desired effects on the operably linked promoter may be revealed. A nucleic acid molecule, such as DNA, is said to be "capable of expressing" a polypeptide if it contains expression control sequences which contain transcriptional regulatory information and such sequences are operably linked to the nucleotide sequence which encodes the polypeptide. For the complete control of gene expression, all transcriptional and translational regulatory elements (or signals) that are operably linked to a heterologous gene should be recognizable by the appropriate host. By "recognizable" in a host is meant that such signals are functional in such host.

The MAdCAM-1 transcriptional regulatory elements of the present invention, obtained through the methods described above, and preferably in a double-stranded form, may be operably linked to a heterologous gene (such as

a reporter gene), preferably in an expression vector, and introduced into a host cell, preferably a eukaryotic cell, to assay reporter gene expression. Preferred eukaryotic cells include choriocarcinoma cell lines, breast cancer cell lines, prostate carcinoma cell lines and kidney cell lines.

As is widely known, translation of eukaryotic mRNA is initiated at the codon that encodes the first methionine. For this reason, it is preferable to ensure that the linkage between a eukaryotic promoter and a reporter gene does not contain any intervening codons that are capable of encoding a methionine. The presence of such codons results either in a formation of a fusion protein (if the AUG codon is in the same reading frame as the DNA encoding the heterologous protein) or a frame-shift mutation (if the AUG codon is not in the same reading frame as the reporter gene).

If desired, a fusion product of a reporter protein may be constructed. For example, the sequence coding for the reporter protein may be linked to a signal sequence which will allow secretion of the protein from, or the compartmentalization of the protein in, a particular host. Such signal sequences may be designed with or without specific protease sites such that the signal peptide sequence is amenable to subsequent removal. Alternatively, the native signal sequence for this protein may be used.

The transcriptional regulatory elements of the invention can be selected to allow for repression or activation, so that expression of the operably linked reporter genes can be modulated. Translational signals are not necessary when it is desired to express antisense RNA sequences or to assay reporter gene expression via mRNA detection.

If desired, the non-transcribed and/or non-translated regions 3′ to the reporter gene can be obtained by the above-described cloning methods. The 3′-non-transcribed region may be retained for its transcriptional termination regulatory sequence elements; the 3'-non-translated region may be retained for its translational termination regulatory sequence elements, or for those elements that direct polyadenylation in eukaryotic cells. Where the native expression

control sequences signals do not function satisfactorily host cell, then sequences functional in the host cell may be substituted.

To transform a mammalian cell with the DNA constructs of the invention many vector systems are available, depending upon whether it is desired to insert the reporter gene product into the host cell chromosomal DNA, or to allow it to exist in an extrachromosomal form. If the reporter gene and an operably linked promoter are introduced into a recipient eukaryotic cell as a non-replicating DNA (or RNA) molecule, which may either be a linear molecule or, more preferably, a closed covalent circular molecule that is incapable of autonomous replication, reporter gene expression may occur through the transient expression of the introduced sequence.

Genetically stable transformants may be constructed with vector systems, or transformation systems, whereby the reporter gene is integrated into the host chromosome. Such integration may occur *de novo* within the cell or, in a most preferred embodiment, be assisted by transformation with a vector that functionally inserts itself into the host chromosome. Vectors capable of chromosomal insertion include, for example, retroviral vectors, transposons or other DNA elements which promote integration of DNA sequences in chromosomes, especially DNA sequence homologous to a desired chromosomal insertion site.

Cells that have stably integrated the introduced DNA into their chromosomes are selected by also introducing one or more markers that allow for selection of host cells which that the desired sequence. For example, the marker may provide biocide resistance, e.g., resistance to antibiotics, or heavy metals, such as copper, or the like. The selectable marker gene can either be directly linked to the reporter gene, or introduced into the same cell by co-transfection. In another embodiment, the introduced sequence is incorporated into a plasmid or viral vector capable of autonomous replication in the recipient host. Any of a wide variety of vectors may be employed for this purpose, as outlined below. Factors of importance in selecting a particular plasmid or viral

vector include: the ease with which recipient cells that contain the vector may be recognized and selected from those recipient cells which do not contain the vector; the number of copies of the vector which are desired in a particular host; and whether it is desirable to be able to "shuttle" the vector between host cells of different species.

Preferred eukaryotic plasmids include those derived from the bovine papilloma virus, vaccinia virus, and SV40. Such plasmids are well known in the art and are commonly or commercially available. For example, mammalian expression vector systems in which it is possible to cotransfect with a helper virus to amplify plasmid copy number, and, integrate the plasmid into the chromosomes of host cells have been described (Perkins, A.S. *et al.*, *Mol. Cell Biol. 3*:1123 (1983); Clontech, Palo Alto, California). Particularly preferred are vectors derived from pCAT-Basic, pCAT-Enhancer and pCAT-Promoter vectors (Promega, Madison, WI).

Once the vector or DNA sequence containing the construct(s) is prepared for expression, the DNA construct(s) is introduced into an appropriate host cell by any of a variety of suitable means, including transfection, electroporation or delivery by liposomes. DEAE dextrin, calcium phosphate, and preferably, the transfection reagent DOTAP, may be useful in the transfection protocol.

After the introduction of the vector *in vitro*, recipient cells are grown in a selective medium, that is, medium that selects for the growth of vector-containing cells. Expression of the reporter gene results in the production mRNA and, if desired, reporter protein. According to the invention, this expression can take place in a continuous manner in the transformed cells, or in a controlled manner. If desired, in *in vitro* culture, the reporter protein is isolated and purified in accordance with conventional conditions, such as extraction, precipitation, chromatography, affinity chromatography, electrophoresis, or the like. Alternatively, levels of reporter protein expression

can be assayed according to conventional protein assays, such as, for example, the CAT expression system.

The MAdCAM-1 transcriptional regulatory elements of the present invention (i.e., the MAdCAM-1 promoter, as well as isolated nucleic acid molecules capable of enhancing and/or repressing gene expression) are useful for screening drugs, ligands and/or other *trans*-acting agents to determine which are capable of affecting expression of MAdCAM-1 or any splice variant thereof. By the invention, *trans*-acting factors can be identified by their ability to up-regulate or down-regulate MAdCAM-1 expression. As used herein, by "MAdCAM-1 *trans*-acting agent" is intended a drug, ligand, or other compound capable interacting, either directly or indirectly, with a MAdCAM-1 transcriptional regulatory element of the present invention to enhance or repress gene expression. Such MAdCAM-1 *trans*-acting elements which interact directly with a transcriptional regulatory element of the present invention include those, which, for example, bind directly to the element and either enhance or repress gene expression. MAdCAM-1 *trans*-acting agents which interact indirectly with a transcriptional regulatory element of the present invention include those which, for example, bind to and induce activity of a second *trans*-acting agent (e.g., a receptor molecule) which itself then, either alone or complexed to the first *trans*-acting agent, binds to the element and either enhances or represses gene expression. One type of *trans*-acting agent is a triplex-forming oligonucleotide. Administration of a suitable oligonucleotide will result in the formation of a triple helix between the oligonucleotide and the MAdCAM-1 promoter, which will inhibit transcription from that promoter (Ebbinghaus, S.W. *et al., Gene Therapy 3*: 287-297 (1996); Roy, C., *Eur. J. Biochem. 220*: 493-503 (1994)). Because the genomic sequence of the region 5' of the MAdCAM-1 gene is given herein (See FIG. 6 and SEQ ID NO:37), one of ordinary skill in the art will readily be able to design suitable oligonucleotides (also called "anti-sense" oligonucleotides) which can inhibit expression from the MAdCAM-1 promoter. One region

which is especially useful for anti-sense design is the 5' untranslated region (*J. Biol. Chem. 266*: 18162-18171 (1991)), which of course is not included in a cDNA, but is included in the genomic sequence disclosed herein.

Thus, in one aspect, the invention provides a screening assay for determining whether any given compound is capable of up-regulating or down-regulating expression from the MAdCAM-1 promoter, leading to an increase or decrease of MAdCAM-1 production.

The screening assay involves (1) providing a host cell transfected with a recombinant nucleic acid molecule containing a MAdCAM-1 transcriptional regulatory element of the present invention and a reporter gene, wherein the transcriptional element is operably linked to the reporter gene; (2) administering a candidate MAdCAM-1 *trans*-acting agent to the transfected host cell; and (3) determining the effect on reporter gene expression.

In a preferred embodiment, the invention provides a screening assay for the identification of substances capable of altering the expression from the MAdCAM-1 promoter, comprising:

(a) measuring the level of expression of a reporter gene in a test cell, wherein said test cell is transformed with a recombinant DNA molecule comprising a reporter gene operably linked to a DNA molecule comprising the promoter of MAdCAM-1, and wherein a candidate MAdCAM-1 *trans*-acting agent is administered to said test cell;

(b) measuring the level of expression of said reporter gene in a control cell, wherein said control cell is transformed with the recombinant DNA molecule of step (a); and

(c) comparing the level of expression of said reporter gene in said test cell to the level of said reporter gene in said control cell.

Suitable and preferred host cells, transfection methods, expression vectors, promoters, and reporter genes, are described above and will be known in the art.

Having generally described the invention, the same will be more readily understood by reference to the following examples, which are provided by way of illustration and are not intended as limiting.

Examples

Example 1: Expression and Purification of any of MAdCAM-1(a-e) in E. coli

The DNA sequence encoding any of the mature MAdCAM-1(a-e) proteins is amplified using PCR oligonucleotide primers specific to the amino terminal sequences of the desired MAdCAM-1(a-e) protein and to vector sequences 3′ to the gene. Additional nucleotides containing restriction sites to facilitate cloning are added to the 5′ and 3′ sequences respectively.

To obtain the DNA sequence encoding MAdCAM-1(a), the plasmid HEBBC23 is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(b), the plasmid HSKCW36 is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(c), the plasmid MAdCAM-1c is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(d), the plasmid MAdCAM-1d is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(e), the plasmid MAdCAM-1e is used, along with the primers given below.

The 5′ oligonucleotide primer has the sequence 5′cgc ccatgg gc cag tcc ctc cag gtg 3′ (SEQ ID NO:11) containing the underlined NcoI restriction site, which encodes 17 nucleotides of the coding sequence of the gene encoding any

of the MAdCAM-1(a-e) proteins shown in FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9), respectively, beginning immediately after the signal peptide.

The 3′ primer has the sequence 5′ cgc aagctt tca ggg cag ctg gtc acc cgc 3′ (SEQ ID NO:12) containing the underlined HindIII restriction site followed by nucleotides complementary to nucleotides 940-967 of FIG. 1, which follow immediately after the coding sequence of any of MAdCAM-1(a-e).

The restriction sites are convenient to restriction enzyme sites in the bacterial expression vector pQE60, which are used for bacterial expression in these examples. (Qiagen, Inc. 9259 Eton Avenue, Chatsworth, CA, 91311). pQE60 encodes ampicillin antibiotic resistance ("Amp^r") and contains a bacterial origin of replication ("ori"), an IPTG inducible promoter, a ribosome binding site ("RBS"), a 6-His tag and restriction enzyme sites.

The amplified DNA encoding any of MAdCAM-1(a-e) and the vector pQE60 both are digested with NcoI and HindIII and the digested DNAs are then ligated together. Insertion of the DNA encoding any of the MAdCAM-1(a-e) proteins into the restricted pQE60 vector places the coding region of MAdCAM-1(a-e) downstream of and operably linked to the vector's IPTG-inducible promoter and in-frame with an initiating AUG appropriately positioned for translation of the appropriate MAdCAM-1(a-e) protein.

The ligation mixture is transformed into competent *E. coli* cells using standard procedures. Such procedures are described in Sambrook et al., MOLECULAR CLONING: A LABORATORY MANUAL, 2nd Ed.; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1989). *E. coli* strain M15/rep4, containing multiple copies of the plasmid pREP4, which expresses lac repressor and confers kanamycin resistance ("Kan^r"), is used in carrying out the illustrative example described herein. This strain, which is only one of many that

are suitable for expressing any of the MAdCAM-1(a-e) proteins, is available commercially from Qiagen.

Transformants are identified by their ability to grow on LB plates in the presence of ampicillin and kanamycin. Plasmid DNA is isolated from resistant colonies and the identity of the cloned DNA confirmed by restriction analysis.

Clones containing the desired constructs are grown overnight ("O/N") in liquid culture in LB media supplemented with both ampicillin (100 μg/ml) and kanamycin (25 μg/ml).

The O/N culture is used to inoculate a large culture, at a dilution of approximately 1:100 to 1:250. The cells are grown to an optical density at 600nm ("OD600") of between 0.4 and 0.6. Isopropyl-B-D-thiogalactopyranoside ("IPTG") is then added to a final concentration of 1 mM to induce transcription from *lac* repressor sensitive promoters, by inactivating the *lac*I repressor. Cells subsequently are incubated further for 3 to 4 hours. Cells then are harvested by centrifugation and disrupted, by standard methods. Inclusion bodies are purified from the disrupted cells using routine collection techniques, and protein is solubilized from the inclusion bodies into 8M urea. The 8M urea solution containing the solubilized protein is passed over a PD-10 column in 2X phosphate-buffered saline ("PBS"), thereby removing the urea, exchanging the buffer and refolding the protein. The protein is purified by a further step of chromatography to remove endotoxin. Then, it is sterile filtered. The sterile filtered protein preparation is stored in 2X PBS at a concentration of 95 μg/ml.

Example 2: Cloning and Expression of any of the MAdCAM-1(a-e) proteins in a Baculovirus Expression System

In this illustrative example, the plasmid shuttle vector pA2 is used to insert the cloned DNA encoding the complete protein, including its naturally associated secretary signal (leader) sequence, into a baculovirus to express any of the mature proteins MAdCAM-1(a-e), using standard methods as described in Summers *et al., A Manual of Methods for Baculovirus Vectors and Insect Cell Culture Procedures,* Texas Agricultural Experimental Station Bulletin No. 1555 (1987). This expression vector contains the strong polyhedrin promoter of the *Autographa californica* nuclear polyhedrosis virus (AcMNPV) followed by convenient restriction sites such as BamHI and Asp718. The polyadenylation site of the simian virus 40 ("SV40") is used for efficient polyadenylation. For easy selection of recombinant virus, the plasmid contains the beta-galactosidase gene from *E. coli* under control of a weak Drosophila promoter in the same orientation, followed by the polyadenylation signal of the polyhedrin gene. The inserted genes are flanked on both sides by viral sequences for cell-mediated homologous recombination with wild-type viral DNA to generate viable virus that express the cloned polynucleotide.

Many other baculovirus vectors could be used in place of the vector above, such as pAc373, pVL941 and pAcIM1, as one skilled in the art would readily appreciate, as long as the construct provides appropriately located signals for transcription, translation, secretion and the like, including a signal peptide and an in-frame AUG as required. Such vectors are described, for instance, in Luckow *et al., Virology 170*:31-39.

The cDNA sequence encoding any of the full length MAdCAM-1(a-e) proteins is amplified using PCR oligonucleotide primers corresponding to the 5′ and 3′ sequences of the gene.

To obtain the DNA sequence encoding MAdCAM-1(a), the plasmid HEBBC23 is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(b), the plasmid HSKCW36 is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(c), the plasmid MAdCAM-1c is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(d), the plasmid MAdCAM-1d is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(e), the plasmid MAdCAM-1e is used, along with the primers given below.

The 5′ primer has the sequence 5′cgc ggatcc gcc atc atg gat ttc gga ctg gcc 3′ (SEQ ID NO:13) containing the underlined BamHI restriction enzyme site followed by 18 bases of the sequence of the relevant MAdCAM-1(a-e) protein shown in FIGS. 1-5, respectively. Inserted into an expression vector, as described below, the 5′ end of the amplified fragment encoding the relevant MAdCAM-1(a-e) protein provides an efficient signal peptide. An efficient signal for initiation of translation in eukaryotic cells, as described by Kozak, M., J. Mol. Biol. 196: 947-950 (1987) is appropriately located in the vector portion of the construct.

The 3′ primer has the sequence 5′cgc ggtacc tca ctt gaa ggg gtc caa gc 3′ (SEQ ID NO:14) containing the underlined Asp718 restriction site followed by nucleotides complementary to nucleotides 1183-1199 of FIG. 1, which follow immediately after the coding sequence of any of MAdCAM-1(a-e).

The cDNA sequence encoding the extracellular soluble domain of any of the MAdCAM-1(a-e) proteins is amplified using PCR oligonucleotide primers corresponding to the 5′ and 3′ sequences of the gene.

To obtain the DNA sequence encoding MAdCAM-1(a), the plasmid HEBBC23 is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(b), the plasmid HSKCW36 is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(c), the plasmid MAdCAM-1c is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(d), the plasmid MAdCAM-1d is used, along with the primers given below.

To obtain the DNA sequence encoding MAdCAM-1(e), the plasmid MAdCAM-1e is used, along with the primers given below.

The 5′ primer has the sequence 5′cgc ggatcc gcc atc atg gat ttc gga ctg gcc 3′ (SEQ ID NO:15), containing the underlined BamHI restriction enzyme site followed by 18 bases of the sequence of the relevant MAdCAM-1(a-e) protein shown in FIGS. 1-5, respectively. Inserted into an expression vector, as described below, the 5′ end of the amplified fragment encoding the relevant MAdCAM-1(a-e) protein provides an efficient signal peptide. An efficient signal for initiation of translation in eukaryotic cells, as described by Kozak, M., J. Mol. Biol. 196: 947-950 (1987) is appropriately located in the vector portion of the construct.

The 3′ primer has the sequence 5′cgc ggtacc tca ggg cag ctg gtc acc cgc 3′ (SEQ ID NO:16) containing the underlined Asp718 restriction site followed by nucleotides complementary to nucleotides 940-967 of FIG. 1, which follow immediately after the coding sequence of any of MAdCAM-1(a-e).

The amplified fragment is isolated from a 1% agarose gel using a commercially available kit ("Geneclean," BIO 101 Inc., La Jolla, Ca.). The fragment then is digested with BamHI and Asp718 and again is purified on a 1% agarose gel. This fragment is designated herein F2.

The plasmid is digested with the restriction enzymes BamHI and Asp718 and then is dephosphorylated using calf intestinal phosphatase, using routine procedures known in the art. The DNA is then isolated from a 1% agarose gel

using a commercially available kit ("Geneclean" BIO 101 Inc., La Jolla, Ca.). This vector DNA is designated herein "V2".

Fragment F2 and the dephosphorylated plasmid V2 are ligated together with T4 DNA ligase. *E. coli* HB101 cells are transformed with ligation mix and spread on culture plates. Bacteria are identified that contain the plasmid with the desired human gene encoding MAdCAM-1(a-e) by digesting DNA from individual colonies using XbaI and then analyzing the digestion product by gel electrophoresis. The sequence of the cloned fragment is confirmed by DNA sequencing. This plasmid is designated herein pBacMAdCAM-1(a-e) (i.e., if MAdCAM-1(a) is cloned, the plasmid is pBacMAdCAM-1(a), while if MAdCAM-1(b) is cloned, the plasmid is pBACMAdCAM-1(b), etc.).

5 μg of the plasmid pBacMAdCAM-1(a-e) is co-transfected with 1.0 μg of a commercially available linearized baculovirus DNA ("BaculoGold™ baculovirus DNA", Pharmingen, San Diego, CA.), using the lipofection method described by Felgner et al., Proc. Natl. Acad. Sci. USA 84: 7413-7417 (1987). 1μg of BaculoGold™ virus DNA and 5 μg of the plasmid pBacMAdCAM-1(a-e) are mixed in a sterile well of a microtiter plate containing 50 μl of serum-free Grace's medium (Life Technologies Inc., Gaithersburg, MD). Afterwards 10 μl Lipofectin plus 90 μl Grace's medium are added, mixed and incubated for 15 minutes at room temperature. Then the transfection mixture is added drop-wise to Sf9 insect cells (ATCC CRL 1711) seeded in a 35 mm tissue culture plate with 1 ml Grace's medium without serum. The plate is rocked back and forth to mix the newly added solution. The plate is then incubated for 5 hours at 27°C. After 5 hours the transfection solution is removed from the plate and 1 ml of Grace's insect medium supplemented with 10% fetal calf serum is added. The plate is put back into an incubator and cultivation is continued at 27°C for four days.

After four days the supernatant is collected and a plaque assay is performed, as described by Summers and Smith, cited above. An agarose gel with "Blue Gal" (Life Technologies Inc., Gaithersburg) is used to allow easy identification and isolation of gal-expressing clones, which produce blue-stained plaques. (A detailed description of a "plaque assay" of this type can also be found in the user's guide for insect cell culture and baculovirology distributed by Life Technologies Inc., Gaithersburg, page 9-10).

Four days after serial dilution, the virus is added to the cells. After appropriate incubation, blue stained plaques are picked with the tip of an Eppendorf pipette. The agar containing the recombinant viruses is then resuspended in an Eppendorf tube containing 200 µl of Grace's medium. The agar is removed by a brief centrifugation and the supernatant containing the recombinant baculovirus is used to infect Sf9 cells seeded in 35 mm dishes. Four days later the supernatants of these culture dishes are harvested and then they are stored at 4°C. A clone containing any of the properly inserted genes encoding MAdCAM-1(a-e) is identified by DNA analysis including restriction mapping and sequencing. This is designated herein as V-MAdCAM-1(a-e), i.e., V-MAdCAM-1(a), or V-MAdCAM-1(b), etc., depending on which MAdCAM-1 variant is cloned.

Sf9 cells are grown in Grace's medium supplemented with 10% heat-inactivated FBS. The cells are infected with the recombinant baculovirus V-MAdCAM-1(a-e) at a multiplicity of infection ("MOI") of about 2 (about 1 to about 3). Six hours later the medium is removed and is replaced with SF900 II medium minus methionine and cysteine (available from Life Technologies Inc., Gaithersburg). 42 hours later, 5 µCi of ^{35}S-methionine and 5 µCi ^{35}S-cysteine (available from Amersham) are added. The cells are further incubated for 16

hours and then they are harvested by centrifugation, lysed and the labeled proteins are visualized by SDS-PAGE and autoradiography.

Example 3: Cloning and Expression in Mammalian Cells

Most of the vectors used for the transient expression of the gene sequence encoding any of MAdCAM-1(a-e) proteins in mammalian cells should carry the SV40 origin of replication. This allows the replication of the vector to high copy numbers in cells (e.g. COS cells) which express the T antigen required for the initiation of viral DNA synthesis. Any other mammalian cell line can also be utilized for this purpose.

A typical mammalian expression vector contains the promoter element, which mediates the initiation of transcription of mRNA, the protein coding sequence, and signals required for the termination of transcription and polyadenylation of the transcript. Additional elements include enhancers, Kozak sequences and intervening sequences flanked by donor and acceptor sites for RNA splicing. Highly efficient transcription can be achieved with the early and late promoters from SV40, the long terminal repeats (LTRs) from Retroviruses, e.g. RSV, HTLVI, HIVI and the early promoter of the cytomegalovirus (CMV). However, cellular signals can also be used (e.g. human actin promoter). Suitable expression vectors for use in practicing the present invention include, for example, vectors such as pSVL and pMSG (Pharmacia, Uppsala, Sweden), pRSVcat (ATCC 37152), pSV2dhfr (ATCC 37146) and pBC12MI (ATCC 67109). Mammalian host cells that could be used include, human Hela, 283, H9 and Jurkart cells, mouse NIH3T3 and C127 cells, Cos 1, Cos 7 and CV1, African green monkey cells, quail QC1-3 cells, mouse L cells and Chinese hamster ovary cells.

Alternatively, the gene can be expressed in stable cell lines that contain the gene integrated into a chromosome. The co-transfection with a selectable marker such as dhfr, gpt, neomycin, hygromycin allows the identification and isolation of the transfected cells.

The transfected gene can also be amplified to express large amounts of the encoded protein. The DHFR (dihydrofolate reductase) is a useful marker to develop cell lines that carry several hundred or even several thousand copies of the gene of interest. Another useful selection marker is the enzyme glutamine synthase (GS) (Murphy *et al., Biochem J. 227*:277-279 (1991); Bebbington *et al., Bio/Technology 10:*169-175 (1992)). Using these markers, the mammalian cells are grown in selective medium and the cells with the highest resistance are selected. These cell lines contain the amplified gene(s) integrated into a chromosome. Chinese hamster ovary (CHO) cells are often used for the production of proteins.

The expression vectors pC1 and pC4 contain the strong promoter (LTR) of the Rous Sarcoma Virus (Cullen *et al., Molecular and Cellular Biology*, 438-4470 (March, 1985)) plus a fragment of the CMV-enhancer (Boshart *et al., Cell 41*:521-530 (1985)). Multiple cloning sites, e.g. with the restriction enzyme cleavage sites BamHI, XbaI and Asp718, facilitate the cloning of the gene of interest. The vectors contain in addition the 3′ intron, the polyadenylation and termination signal of the rat preproinsulin gene.

Example 3(a): Cloning and Expression in COS Cells

The expression plasmid, pMAdCAM-1(a-e) HA, is made by cloning a cDNA encoding one of MAdCAM-1(a-e) into the expression vector pcDNAI/Amp (which can be obtained from Invitrogen, Inc.).

The expression vector pcDNAI/amp contains: (1) an *E.coli* origin of replication effective for propagation in *E. coli* and other prokaryotic cells; (2) an ampicillin resistance gene for selection of plasmid-containing prokaryotic cells; (3) an SV40 origin of replication for propagation in eukaryotic cells; (4) a CMV promoter, a polylinker, an SV40 intron, and a polyadenylation signal arranged so that a cDNA conveniently can be placed under expression control of the CMV promoter and operably linked to the SV40 intron and the polyadenylation signal by means of restriction sites in the polylinker.

A DNA fragment encoding the relevant MAdCAM-1(a-e) protein is cloned into the polylinker region of the vector so that recombinant protein expression is directed by the CMV promoter. The plasmid construction strategy is as follows. The cDNA encoding the relevant MAdCAM-1(a-e) is amplified using primers that contain convenient restriction sites, much as described above regarding the construction of expression vectors for expression of the desired MAdCAM-1(a-e) in *E. coli*.

Suitable primers include the following, which are used in this example.

The DNA sequence encoding the full length protein of any of MAdCAM-1(a-e) is amplified using PCR oligonucleotide primers corresponding to the 5' and 3' sequences of the gene:

The 5' primer has the sequence 5' cgc ggatcc gcc atc atg gat ttc gga ctg gcc 3' (SEQ ID NO:17) containing the underlined BamH1 restriction enzyme site followed by 18 bases of the sequence of the relevant MAdCAM-1(a-e) gene shown in FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9), respectively. Inserted into an expression vector, as described below, the 5' end of the amplified fragment encoding any of human MAdCAM-1(a-e) provides an efficient signal peptide. An efficient signal for initiation of translation in eukaryotic cells, as described by

Kozak, M., J. Mol. Biol. 196:947-950 (1987) is appropriately located in the vector portion of the construct.

The 3' primer has the sequence 5' cgc ggtacc tca ctt gaa ggg gtc caa gc 3' (SEQ ID NO:18) containing the underlined Asp718 restriction followed by nucleotides complementary to nucleotides 1183-1199 of the MAdCAM-1(a) coding sequence given in FIG. 1.

In order to clone a gene encoding the extracellular soluble domain of MAdCAM-1(a-e), the 5′ primer, containing the underlined BamHI site, an AUG start codon and 18 codons of the 5′ coding region has the following sequence:

5′ cgc ggatcc gcc atc atg gat ttc gga ctg gcc 3′ (SEQ ID NO:19).

The 3′ primer, containing an XbaI site, a stop codon, and 3′ coding sequence for the extracellular domain, has the following sequence:

5′ cgc tctaga tca agc gta gtc tcc gac gtc gta tgg gta 3′ (SEQ ID NO:20).

The PCR amplified DNA fragment and the vector, pcDNAI/Amp, are digested with HindIII and XhoI and then ligated. The ligation mixture is transformed into *E. coli* strain SURE (available from Stratagene Cloning Systems, 11099 North Torrey Pines Road, La Jolla, CA 92037), and the transformed culture is plated on ampicillin media plates which then are incubated to allow growth of ampicillin resistant colonies. Plasmid DNA is isolated from resistant colonies and examined by restriction analysis and gel sizing for the presence of a fragment encoding the relevant MAdCAM-1(a-e).

For expression of recombinant MAdCAM-1(a-e), COS cells are transfected with an expression vector, as described above, using DEAE-DEXTRAN, as described, for instance, in Sambrook et al., MOLECULAR CLONING: A LABORATORY MANUAL, Cold Spring Laboratory Press, Cold

Spring Harbor, New York (1989). Cells are incubated under conditions for expression of MAdCAM-1(a-e) by the vector.

Expression of the MAdCAM-1(a-e)HA fusion protein is detected by radiolabelling and immunoprecipitation, using methods described in, for example Harlow et al., ANTIBODIES: A LABORATORY MANUAL, 2nd Ed.; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1988). To this end, two days after transfection, the cells are labeled by incubation in media containing ^{35}S-cysteine for 8 hours. The cells and the media are collected, and the cells are washed and the lysed with detergent-containing RIPA buffer: 150 mM NaCl, 1% NP-40, 0.1% SDS, 1% NP-40, 0.5% DOC, 50 mM TRIS, pH 7.5, as described by Wilson et al. cited above. Proteins are precipitated from the cell lysate and from the culture media using an HA-specific monoclonal antibody. The precipitated proteins then are analyzed by SDS-PAGE gels and autoradiography. An expression product of the expected size is seen in the cell lysate, which is not seen in negative controls.

Example 3(b): Cloning and Expression in CHO Cells

The vector pC1 is used for the expression of any of the MAdCAM-1(a-e) proteins. Plasmid pC1 is a derivative of the plasmid pSV2-dhfr [ATCC Accession No. 37146]. Both plasmids contain the mouse DHFR gene under control of the SV40 early promoter. Chinese hamster ovary- or other cells lacking dihydrofolate activity that are transfected with these plasmids can be selected by growing the cells in a selective medium (alpha minus MEM, Life Technologies) supplemented with the chemotherapeutic agent methotrexate. The amplification of the DHFR genes in cells resistant to methotrexate (MTX) has been well documented (see, e.g., Alt, F.W., Kellems, R.M., Bertino, J.R., and

Schimke, R.T., 1978, J. Biol. Chem. 253:1357-1370, Hamlin, J.L. and Ma, C. 1990, Biochem. et Biophys. Acta, 1097:107-143, Page, M.J. and Sydenham, M.A. 1991, Biotechnology Vol. 9:64-68). Cells grown in increasing concentrations of MTX develop resistance to the drug by overproducing the target enzyme, DHFR, as a result of amplification of the DHFR gene. If a second gene is linked to the DHFR gene it is usually co-amplified and over-expressed. It is state of the art to develop cell lines carrying more than 1,000 copies of the genes. Subsequently, when the methotrexate is withdrawn, cell lines contain the amplified gene integrated into the chromosome(s).

Plasmid pC1 contains for the expression of the gene of interest a strong promoter of the long terminal repeat (LTR) of the Rous Sarcoma Virus (Cullen, et al., Molecular and Cellular biology, March 1985, 438-4470) plus a fragment isolated from the enhancer of the immediate early gene of human cytomegalovirus (CMV) (Boshart et al., Cell 41:521-530, 1985). Downstream of the promoter is a BamHI restriction enzyme cleavage site that allows the integration of the genes. Behind this cloning site the plasmid contains translational stop codons in all three reading frames followed by the 3' intron and the polyadenylation site of the rat preproinsulin gene. Other high efficient promoters can also be used for the expression, e.g., the human β-actin promoter, the SV40 early or late promoters or the long terminal repeats from other retroviruses, e.g., HIV and HTLVI. For the polyadenylation of the mRNA other signals, e.g., from the human growth hormone or globin genes can be used as well.

Stable cell lines carrying a gene of interest integrated into the chromosomes can also be selected upon co-transfection with a selectable marker such as gpt, G418 or hygromycin. It is advantageous to use more than one selectable marker in the beginning, e.g. G418 plus methotrexate.

The plasmid pC1 is digested with the restriction enzyme BamHI and then dephosphorylated using calf intestinal phosphates by procedures known in the art. The vector is then isolated from a 1% agarose gel.

The DNA sequence encoding the full length protein of any of MAdCAM-1(a-e) is amplified using PCR oligonucleotide primers corresponding to the 5' and 3' sequences of the gene:

The 5' primer has the sequence 5' cgc ggatcc gcc atc atg gat ttc gga ctg gcc 3' (SEQ ID NO:17) containing the underlined BamH1 restriction enzyme site followed by 18 bases of the sequence of the relevant MAdCAM-1(a-e) gene shown in FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9), respectively. Inserted into an expression vector, as described below, the 5' end of the amplified fragment encoding any of human MAdCAM-1(a-e) provides an efficient signal peptide. An efficient signal for initiation of translation in eukaryotic cells, as described by Kozak, M., J. Mol. Biol. 196:947-950 (1987) is appropriately located in the vector portion of the construct.

The 3' primer has the sequence 5' cgc ggtacc tca ctt gaa ggg gtc caa gc 3' (SEQ ID NO:18) containing the underlined Asp718 restriction followed by nucleotides complementary to nucleotides 1183-1199 of the MAdCAM-1(a) coding sequence given in FIG. 1.

The DNA sequence encoding the extracellular soluble domain of any of MAdCAM-1(a-e) proteins is amplified using PCR oligonucleotide primers corresponding to the 5' and 3' sequences of the gene:

The 5' primer has the sequence 5' cgc ggatcc gcc atc atg gat ttc gga ctg gcc 3' (SEQ ID NO:17) containing the underlined BamH1 restriction enzyme site followed by 18 bases of the sequence of the relevant MAdCAM-1(a-e) gene shown in FIGS. 1-5 (SEQ ID NOs:1, 3, 5, 7, 9), respectively. Inserted into an expression vector, as described below, the 5' end of the amplified fragment

encoding any of human MAdCAM-1(a-e) provides an efficient signal peptide. An efficient signal for initiation of translation in eukaryotic cells, as described by Kozak, M., J. Mol. Biol. 196:947-950 (1987) is appropriately located in the vector portion of the construct.

The 3' primer has the sequence 5' cgc ggtacc tca ggg cag ctg gtc acc cgc 3' (SEQ ID NO:21) containing the underlined Asp718 restriction followed by nucleotides complementary to nucleotides 940-967 of the MAdCAM-1(a) coding sequence given in FIG. 1.

The amplified fragments are isolated from a 1% agarose gel as described above and then digested with the endonuclease BamHI and then purified again on a 1% agarose gel.

The isolated fragment and the dephosphorylated vector are then ligated with T4 DNA ligase. E.coli HB101 cells are then transformed and bacteria identified that contained the plasmid pC1 inserted in the correct orientation using the restriction enzyme BamHI. The sequence of the inserted gene is confirmed by DNA sequencing.

Transfection of CHO-DHFR-cells

Chinese hamster ovary cells lacking an active DHFR enzyme are used for transfection. 5 μg of the expression plasmid C1 are cotransfected with 0.5 μg of the plasmid pSVneo using the lipofecting method (Felgner et al., supra). The plasmid pSV2-neo contains a dominant selectable marker, the gene neo from Tn5 encoding an enzyme that confers resistance to a group of antibiotics including G418. The cells are seeded in alpha minus MEM supplemented with 1 mg/ml G418. After 2 days, the cells are trypsinized and seeded in hybridoma cloning plates (Greiner, Germany) and cultivated from 10-14 days. After this period,

single clones are trypsinized and then seeded in 6-well petri dishes using different concentrations of methotrexate (25 nM, 50 nM, 100 nM, 200 nM, 400 nM). Clones growing at the highest concentrations of methotrexate are then transferred to new 6-well plates containing even higher concentrations of methotrexate (500 nM, 1 μM, 2 μM, 5 μM). The same procedure is repeated until clones grow at a concentration of 100 μM.

The expression of the desired gene product is analyzed by Western blot analysis and SDS-PAGE.

Example 4: Tissue distribution of expression of MAdCAM-1(a-e) proteins

Northern blot analysis was carried out to examine expression of the MAdCAM-1(a) gene in human tissues, using methods described by, among others, Sambrook *et al.*, cited above. A cDNA probe containing the entire nucleotide sequence of the gene encoding the MAdCAM-1(a) protein (SEQ ID NO:1) was labeled with ^{32}P using the *redi*prime™ DNA labeling system (Amersham Life Science), according to manufacturer's instructions. After labeling, the probe was purified using a CHROMA SPIN-100™ column (Clontech Laboratories, Inc.), according to manufacturer's protocol number PT1200-1. The purified labeled probe was then used to examine various human tissues for mRNA corresponding to any of MAdCAM-1(a).

Multiple Tissue Northern (MTN) blots containing various human tissues (H) or human immune system tissues (IM) were obtained from Clontech and were examined with labeled probe using ExpressHyb™ hybridization solution (Clontech) according to manufacturer's protocol number PT1190-1. Following hybridization and washing, the blots were mounted and exposed to film at -70°C overnight, and films developed according to standard procedures.

The blots revealed that MAdCAM-1(a) is expressed strongly in small intestine, less strongly in colon and spleen, and very weakly in pancreas and brain.

Example 5: Sequence Analysis of Human MAdCAM-1 cDNAs and Genomic Clones

Materials and Methods

Isolation of human MAdCAM-1 cDNA and genomic clones

Human MAdCAM-1 cDNA was initially identified as an expressed sequence tag (EST) following screens for homology in an EST cDNA database (Adams, M.D., *et al. Nature 377*:3-17 (1995); Adams, M.D. *et al. Science 252*:1651-1656 (1991); Adams, M.D., *et al. Nature 355*: 632-63444 (1992)) using the BLAST network service provided by the National Center for Biotechnology Information. Partial-length MAdCAM-1 cDNA clones HEBBC23X and HEBBC23Y were identified in a database from an early stage human brain cDNA library. The library was constructed as described previously (Adams, M.D., *et al. Nature 377*:3-17 (1995)) using the Lambda ZAP II vector (Stratagene, La Jolla, California) from cDNA synthesized according to the method of Gubler and Hoffman. A MAdCAM-1 genomic clone was subsequently isolated by screening a cosmid library constructed in the cosmid vector pCV007 (Choo, K. H., *et al., Gene 46*: 277 (1986)). The library was replica plated onto Gene-Screen Plus filters (DuPont, Boston, MA), and screened as described previously (Leung, E., *et al. Int. Immunol. 5*: 551-558 (1993)) with the insert of the MAdCAM-1 EST clone labeled by random hexanucleotide priming (see Example 6).

DNA sequencing

DNA sequences were determined by cycle sequencing using Applied Biosystems automated DNA sequenators (The Centre for Gene Technology, School of Biological Sciences, University of Auckland, Auckland, New Zealand; and at Human Genome Sciences Inc., Rockville, Maryland). The complete composite MAdCAM-1 sequencer obtained from genomic and cDNA clones was determined on both strands using a combination of universal M13 primers, and primers specific for human MAdCAM-1 sequences. A MAdCAM-1 genomic clone was subsequently isolated by screening a cosmid library constructed in the cosmid vector pCV007 (Choo, K. H., *et al., Gene 46*: 277 (1986)). The library was replica plated onto Gene-Screen Plus filters (DuPont, Boston, MA), and screened as described previously (Leung, E., *et al. Int. Immunol. 5*: 551-558 (1993)) with the insert of the MAdCAM-1 EST clone labeled by random hexanucleotide priming.

PCR amplification and identification of MAdCAM-1 splice variants

For PCR amplification to detect MAdCAM-1 variants, ten micrograms of total RNA from human fetal brain (Clontech) in reverse transcriptase (RT) buffer (BRL, Gaithersburg, MD) was heated to 70°C for 3 min and then cooled on ice. All four dNTPs were added to a final concentration of 0.5 mM, together with 500 ng of random hexamer primers, and 400 U of Superscript RT (BRL, Life Technologies Inc. MD, USA) in a total volume of 20 µl. The random priming reaction was incubated at 42°C for 2 h. Two ml of this cDNA was subjected to 20 cycle of amplification in a thermocycler (95°C 30 sec; 63°C 30 sec; 72°C 30 sec) with 100 ng primer U166+ (SEQ ID NO:22) (5′-CGC TCT

CCT TCT CCC TGC TC-3′) and 100 ng of primer L776- (SEQ ID NO:23) (5′TGG TGG GTG GGT GTC GTC CTC A-3′), using a final dNTP concentration of 200 μM and 2.6 U of Expand (Boehringer Mannheim). The U166+ and L776 primers correspond to the sequences 435-454 and 1047-1068 of human MAdCAM-1. An aliquot of 2.5 μl of the PCR reaction was reamplified For 25 cycles using the U166+ primer, and the nested primer L743- (SEQ ID NO:24) 5′-CGG CAG CGT TTC CAG AGG TGA TAC-3′) corresponding to nucleotides 1013-1037, with the same annealing temperature. The PCR product was ethanol precipitated and ligated into an EcoRV digested, Taq polymerase 3′ dTTP-tailed pBluescript vector, and sequenced. PCR was also used as described above to demonstrate continuity between genomic MAdCAM-1 5′-sequences and the MAdCAM-1 EST. Twenty cycles of amplification were carried out (95°C 30 sec; 69°C 45 sec; 72°C 45 sec) with 100 ng primer U203 (SEQ ID NO:25) (5′-GGGACTGAGCATGGATTT CGACTGGCCCT-3′) and 100 ng of primer L103 (SEQ ID NO:26) (5′CGTACAGGCCACCTCCGGGTCACCAGGCA-CCA-3′), using a final dNTP concentration of 200 μM and 2.6 U of Expand (Boehringer Mannheim). The L103 primer corresponds to the sequence 347-405 of the human MAdCAM-1. An aliquot of 2.5 μl of the PCR reaction was reamplified for 25 cycles using the L203 primer, and the nested primer L50- (SEQ ID NO:27) (5′-GCTGGT CCGGGAAGGCGTACACAA GGAGCTGC-3′) corresponding to nucleotides 321-352, with the same annealing temperature. The PCR product was ethanol precipitated and ligated into an EcoRV digested, Taq polymerase 3′ dTTP-tailed pBluescript vector, and sequenced.

Northern blot analysis

For northern analysis, MTN (Clontech) filters were screened with the insert of the MAdCAM-1 EST clone labeled by random hexanucleotide priming. The conditions of hybridization were 1% SDS, 2 x SSC, 10% (w/v) dextran sulphate, 100 μg/ml denatured salmon sperm DNA, and 50% (v/v) deionized formamide at 50°C. Filters were washed twice in 0.1 x SSC, 0.1% SDS at 50°C for 30 min. and autoradiographed using XAR-5 film and Cronex Lightning Plus screens.

Results and discussion

A database of human ESTs was searched for homologs of mouse MAdCAM-1 by using the BLAST algorithm (Altschul, S. F., *et al. J. Mol. Biol. 215*:403-410 (1990)). Partial overlapping MAdCAM-1 cDNA clones HEBBC23X and HEBBC23Y were initially identified from an early stage human brain cDNA library (Figure 8A). They were sequenced on both strands and together encoded the MAdCAM-1 sequence from a position corresponding to amino acid residue 89 of the mouse MAdCAM-1 cDNA clone pMAd-7, to the end of the 3′-untranslated region. HEBBC23Y and X encoded from nucleotide positions 273 to 858, and 544 to 1536, respectively of the human MAdCAM-1 sequence. In order to obtain the missing 5′-end sequence, the early stage brain library was rescreened, as well as five other brain, pancreatic, and adult and fetal spleen cDNA libraries, but no clones that extended the sequence were obtained. As an alternative approach, fetal brain mRNA was subjected to rapid amplification of cDNA ends (RACE), but despite exhaustive attempts the MAdCAM-1 5′-sequence remained elusive. As a last resort 100,000 colonies of

a genomic library in the cosmid vector, pCV007, (Choo, K. H., *et al., Gene 46*: 277 (1986)) were screened with the MAdCAM-1 EST cDNA clone (see Example 6). Of several clones isolated, one strongly hybridizing clone, MAD-C1, was characterized and found to contain the missing sequence on a 5 kb Sac I-Sac I fragment. Continuity between the cosmid and cDNA sequences was established by RT-PCR from fetal brain RNA using a sense primer U203 to putative genomic MAdCAM-1 5′-untranslated and signal peptide sequence, and nested antisense primers L50 and L103 to the 5′-end of the EST clone (see Methods and Example 6).

The composite nucleotide and deduced amino acid sequences of the MAdCAM-1 HEBBC23X cDNA clone, the genomic clone MAD-C1, and the 5′-PCR product are given in Fig. 8. The nucleotide sequence of 1546 bp ends with the polyadenylation signal AAATAAA (SEQ ID NO:28), followed 15 bases further by a poly(A) stretch. Ten bp of the 5′-untranslated sequence has been added for completeness. The open reading frame beginning with an ATG at position 1 encodes a protein of 382 amino acid residues. The ATG start codon, which is flanked by the consensus sequence Pur XXAUG Pur (SEQ ID NO:29), is followed by a predominantly hydrophobic segment of 18 amino acid residues characteristic of a signal peptide. A hydropathicity plot of the deduced amino acid sequence (Fig. 7) revealed a sequence presumed to be the transmembrane domain, encompassing residues 320 to 339. Thus, the sequence predicts a transmembrane bound protein comprised of a predominantly hydrophilic 103 amino acid extracellular domain, a 20 amino acid transmembrane segment, and a 43 amino acid cytoplasmic domain, with an Mr of 38,340. There is a single potential N-linked glycosylation site at amino acid position 83.

The deduced amino acid sequence revealed a 17 amino acid signal peptide, two immunoglobulin (Ig)-line domains, an 86 amino acid mucin-like

region rich in serine/threonine residues, a 20 amino acid transmembrane domain, and a 43 amino acid charged cytoplasmic domain. The sequences of the two N-terminal Ig-like domains are highly conserved (59-65%) with the corresponding receptor-binding Ig domains of mouse MAdCAM-1. No counterpart to the third IgA-like domain of mouse MAdCAM-1 was present, and instead the serine/threonine-rich mucin domain has been extended as two distinguishable regions, here designated the major and minor mucin domains. The major domain is formed from six tandem repeats of an eight amino acid sequence having the consensus DTTSPEP/SP (SEQ ID NO:30), which is similar to the imperfect repeats of the intestinal mucin MUC-2. The mucin domains of the MAdCAM-1 human/mouse species homologs are distinct, in accord with the notion that mucin domains are not phylogenetically conserved. Human MAdCAM-1 mRNA transcripts were restricted to small intestine, colon, spleen, pancreas, and brain which is a further indication that the clones encode MAdCAM-1. Alternatively spliced MAdCAM-1 variants were identified that lack all or part of the second Ig domain, and all or part of the major mucin domain, indicating that the function of this vascular addressin might be regulated by extensive modifications to its multidomain structure.

The extracellular domain comprises two Ig-like domains of 52 and 69 amino acid residues, respectively, each possessing the invariant cysteine residues that stabilize the immunoglobulin loop; with the first domain having doublet cysteines. There is a mucin-like 48 amino acid residue domain encompassing residues 226-273, which is rich in serine, threonine, and proline residues (71%). The mucin domain is formed from six tandem repeats of an eight amino acid sequence having the general consensus DTTSPEP/SP (SEQ ID NO:30). The repeats are highly conserved with one another (75-100%), suggesting that they arose by duplication. This domain has 19 potential sites for O-linked

glycosylation. The mucin-like nature of the region extends to a lesser degree as far as the transmembrane domain, since the serine/threonine/proline content is still quite high (43%). We designate this latter region (positions 278 to 311) as the minor mucin domain, and the mucin tandem repeats immediately 5′ as the major mucin domain. A search of the NBRF database revealed that human MAdCAM-1 was most similar to mouse MAdCAM-1, but striking homologies were also identified with VCAM-1, and ICAM-1. Alignment of the human and mouse sequences (not shown) revealed an overall weak similarity of 39%. However, Ig domains 1 and 2 in particular have been highly conserved, 59 and 65%, respectively; and similarity increases to 69 and 81%, respectively, when conservative substitutions are included. This is to be expected since these two Ig domains interact to support binding to the LPAM-1 receptor, and both domains are required for full function. The membrane-proximal regions of the extracellular domains of human and mouse MAdCAM-1 are peptide backbones designed for decoration with complex O-linked carbohydrate moieties for presentation to L-selectin, and as such, only the serine/threonine/proline content needs to be conserved. Hence, after the first mucin repeat there is little similarity between the human and mouse sequences, except for transmembrane domain which is 55% identical. The short charged cytoplasmic domains share only 35% identity, and the human sequence extends 24 amino acid residues further than the mouse sequence. Clone HEBBC23X lacks an equivalent of the third Ig domain of mouse MAdCAM-1. A truncated mouse MAdCAM-1 variant has been identified in which exon 4 is spliced out removing both the mucin domain and the third Ig domain (Sampaio *et al., J. Immunol. 155*: 2477-86 (1995)). The third Ig domain of mouse MAdCAM-1 is strikingly similar to the Cα2 constant region immunoglobulin loop of human and gorilla IgA1 (Briskin *et al., Nature 363*:461-64 (1993)). It was suggested that it may be able to interact with IgA-specific Fcα

receptors or related surface receptors on mucosal T cells, given that the Cα2 constant regions mediates IgA interactions with the poly-immunoglobulin Fc receptor. It remains plausible that an Ig domain with a mucosal function is not needed in the human brain, and that a human counterpart to the three Ig domain form cloned from the mouse hiMAd-4 brain endothelioma cell line might be expressed in human PP or mesenteric venules. Completion of the sequence analysis of the MAdCAM-1 cosmid clone should resolve this point.

Human MAdCAM-1 may have compensated for a lack of a third Ig domain by having two mucin domains to hold the two N-terminal ligand-binding domains above the glycocalyx for presentation to LPAM-1. In mouse there are 108 amino acid residues separating the mucin domain from the transmembrane domain compared to only 46 residues separating the major mucin domain from the transmembrane domain in human MAdCAM-1. The distances may not be so dissimilar given that the third Ig domain of mouse MAdCAM-1 is a loop structure, whereas the extended mucin domain in human MAdCAM-1 is probably rod-like as are the mucin repeats of MUC-1. (Fontenot *et al., Cancer Res. 53*: 5386-94 (1993)). The repeats in the major mucin domain may have been inserted, possibly by a gene conversion event involving a mucin gene, to enrich the overall content of serine/threonine residues (40% in major domain) and to enable better presentation to L-selectin by positioning the major mucin repeat above the glycocalyx.

A search of the NBRF database with the sequence of the tandem repeats of the major mucin domain revealed most similarity (up to 62% including conservative substitutions) with a region of imperfect repeats in the human intestinal mucin MUC-2. MUC-2 contains two distinct regions with a high degree of internal homology. (Toribara *et al., J. Clin. Invest. 88*: 1005-13 (1991)). There is a region of imperfect repeats that range from 7 to 40 amino

acids, with the most common length being 16 amino acids. This 385 amino acid region has a high threonine (47.8%), proline (35.6%) and serine (10.6%) content. It is this region to which MAdCAM-1 shares similarity (Fig. 2). The major MAdCAM-1 tandem repeat domain is not as rich in such residues, and 22% of the dissimilar amino acids are acidic residues which are totally absent from the imperfect repeats of MUC-2. In MUC-2 there is also a 3' region composed of 69 bp tandem repeats arranged in an array of up to 115 units, which is not similar to the MAdCAM-1 mucin region despite having a high serine/threonine/proline content (87%). (Zrihan-licht *et al., Eur J. Biochem 224*: 787-95 (1994)). The human intestinal mucin MUC-1 has a serine/threonine/proline-rich 20 amino acid residue domain (SEQ ID NO:31) PDTRPAPGSTAPPAHGVTSA, repeated up to 200 times, (Gum *et al., J. Biol. Chem.* 266: 22733-38 (1991)) and rat intestinal mucin has the repeat sequence (SEQ ID NO:32) TTTPDV, (Spicer *et al., J. Biol. Chem. 266*: 15099-109 (1991)) but neither of these sequences bear similarly to MAdCAM-1. Repetitive portions of intestinal mucin genes are not well conserved phylogenetically, and this may explain the divergence of the human and mouse MAdCAM-1 3' sequences. (Vos *et al., Biochem. Biophys. Res. Commun. 181*: 121-30 (1991); Shimizu & Shaw, *Nature 366*: 630-31 (1993)). Thus the primary function of the MAdCAM-1 mucin repeats is probably purely to provide a framework for extensive O-linked glycosylation.

MAdCAM-1 clone HEBBC23Y appears to be a splice variant in the 3 mucin repeats are missing (amino acid residues 231-254) (Figs. 8A, 10). In order to determine whether additional mucin domain splice variants might exist, MAdCAM-1 transcripts were amplified from human fetal brain using sense and antisense PCR primers designed to the start of Ig domain 2 and the cytoplasmic domain, respectively. Several novel splice variants were identified including one which lacked almost all of the second Ig domain and all the major mucin repeats;

and two others which had lost half of Ig domain 2 and 2 to 3 mucin repeats (Fig. 10A). Several of these alternatively spliced transcripts could be accommodated in the broad band seen on northerns, whereas those with larger deletions may be more weakly expressed and visible as a faint leader smear, as is the case for the alternatively spliced variant of mouse MAdCAM-1 (Sampaio *et al., J. Immunol. 155*: 2477-86 (1995)). None of the splice sites correlate with exon/intron boundaries identified in the mouse MAdCAM-1 gene, and hence they probably represent internal splice donor and acceptor sites within the respective exons (Fig. 10B). Alternative splicing of human MAdCAM-1 is in accord with alternative splicing of its mouse homologue. A proposed single Ig domain form (Fig. 11) containing just Ig domain 1 is interesting since analysis of the structural requirements for mouse MADCAM-1 ligand-binding revealed both N-terminal Ig domains were required for full function. Nevertheless a mouse MAdCAM-1 chimeric molecule lacking Ig domain 2 could bind to LPAM-1 (to a lesser extent), but only after integrin activation. The proposed naturally occurring Ig domain 2-deficient form of MAdCAM-1, identified in this report, may prove to be specialized to be more sensitive to the activation/inactivation status of LPAM-1.

The regulation of mucin adhesion by alternative splicing is well established. (Thomas, M. L., *Ann Rev. Immunol. 7*: 339-69 (1989)). The leukocyte common antigen family, for instance, is generated by alternative splicing of three exons encoding a mucin-like region. (Berg *et al., Cellular and Molecular Mechanisms of Inflammation* 2:111-29 (1991)). The MAdCAM-1 variants described in this report possess from 0 to 6 mucin repeats (Fig. 11), and might be expected to vary in their affinity for L-selectin. Whether there is a spatio/temporal patterns or stochastic expression of alternatively spliced forms of MAdCAM-1 on the surface of venules remains to be determined.

Multiple Human Tissue mRNA Northern blots (MTN and MTNII, Clontech) were probed with the cDNA clone HEBBC23X revealing a transcript of 1.6 kb, expressed very strongly in small intestine, less strongly in colon and spleen, and very weakly in pancreas and brain. These results are consistent with northern and immunohistological studies of mouse MAdCAM-1, which revealed expression in PP, MLN, at low levels in PLN, and some expression in the marginal sinus around splenic white pulp nodules in the spleen. (Briskin *et al.*, *Nature 363*:461-64 (1993); Kraal *et al., Am. J. path. 147*: 763-771 (1995).

In summary, several features of mouse MAdCAM-1 have been stringently conserved in humans. This includes the tissue distribution of human MAdCAM-1, and the structure of the two Ig ligand-binding domains; yet the 3'-region is quite divergent. In accord with the regulation of other mucins and IgCAMs, the function of human MAdCAM-1 is likely to be regulated by extensive alternative splicing as evidenced by the variant forms described herein.

Example 6: Genomic Organization and Mapping of the Human MAdCAM-1 Gene; Analysis of the 5'-Promoter Region

Materials and Methods

Isolation of MAdCAM-1 cosmid and genomic phage clones

The two human genomic libraries screened were a Stratagene l FIX II library prepared from human placenta genomic DNA digested with MboI, and a cosmid library constructed in the vector pAVCV007 from DNA partially digested with Mbo I. The cosmid library was replica plated onto Gene-Screen Plus filters (Du Pont, Boston, MA), and screened with the Xho I-EcoR I 32P-labeled 500 bp insert of the MAdCAM-1 cDNA clone PCR Y.

Positive clones HEBBC23591 and GM3 were isolated from the phage and cosmid libraries, respectively.

Subcloning of restriction fragments and sequence determination

Restriction enzyme fragments of the genomic clones were subcloned into pBluescript and sequenced with a panel of oligonucleotide primers designed to the MAdCAM-1 cDNA sequence. DNA sequence was determined by cycle sequencing using an Applied Biosystems 373A automated DNA sequencer (School of Biological Sciences, University of Auckland, Auckland). The entire transcribed regions of the MAdCAM-1 gene, previously defined by the MAdCAM-1 cDNA, were identified and sequenced. Exon-intron boundaries were assigned by direct comparison of the cDNA and genomic sequences, and according to the GT/AG rule for splicing. The determined DNA sequence has been submitted to GenBank databank.

Chromosomal mapping of the human MAdCAM-1 gene

A combination of PCR analysis of a panel of human/rodent somatic cell hybrids and fluorescence in situ hybridization (FISH) to human metaphase chromosomes was used to define the chromosomal location of the MAdCAM-1 gene. Fourteen of the cell hybrids contained a single human chromosome, whereas the remaining 10 contained 2 to 3 chromosomes, or 1 to 3 chromosomal fragments. Two primers U707 and L1072 were designed to nucleotide positions 978-999 (SEQ ID NO:39) (TGC GGT GCT GGG ACT GCT GCT C, sense) and 1344-1364 (SEQ ID NO:40) (TCA GGG AGG GGC TTC AGG TCA, antisense) of the MAdCAM-1 cDNA sequence, respectively. They amplified a PCR

product of 386 bp from human DNA, but not from mouse or hamster DNA. The PCR conditions were: 5 min at 95°C, followed by 30 cycles of 94°C for 30 s, 60°C for 30 s, 72°C for 30 s, and a final extension at 72°C for 5 min.

The precise regional localization of the MAdCAM-1 gene was determined by single copy gene fluorescence *in situ* hybridization (FISH) to human male metaphase chromosome spreads. Briefly, a 1.3 kb MAdCAM-1 cDNA was nick-translated using digoxygenin 11-dUTP (Boehringer Mannheim), and FISH was carried out. Individual chromosomes were counterstained with 4'-6-diamidino-2-phenyindole-2HCl (DAPI). Color digital images containing both DAPI bands and gene signal detected with anti-digoxygenin-tagged rhodamine fluorescent label were recorded using a triple-band pass filter set (Chroma Technology, Inc., Brattleburo, VT) in combination with a charged coupled-device camera (Photometrics, Inc., Tucson, AZ) and variable excitation wave length filters. Images were analyzed using the ISEE software package (Inovision Corp., Durham, N.C.).

Construction of human MAdCAM-1-luciferase fusion genes for assays of promoter activity

A 700 base pair fragment encoding a region immediately 5' of the MAdCAM-1 gene and including the translational start site was PCR amplified from a Sac I-Pst I subclone of the cosmid clone pGM3 using the T7 forward primer (SEQ ID NO:41) (5'-GTA ATA CGA CTC ACT ATA GG-3'; sense) and the MAdCAM-1-specific antisense primer MAD-2 (SEQ ID NO:42) (5'-AGG GCC AGT CCG AAA TCC ATG CTC AGT CCC-3'). The PCR product was subcloned into the EcoRV site in pBluescript, excised with Hind III and subcloned into the pGL-2 Basic vector (Promega, Madison, WI) which contains

a firefly luciferase reporter gene. The insert of the clone created, pGL-2/B-718, was sequenced, confirming that no PCR errors had been incorporated.

Genomic organisation of the human MAdCAM-1 gene

In order to isolate the MAdCAM-1 gene, 200,000 colonies of a genomic library in the cosmid vector, pAVCV007, were screened with the MAdCAM-1 cDNA clone PCR Y that encodes from nucleotide positions 273 to 858. Of two clones obtained, the longest, GM3, contained the entire gene, and 5'-untranslated region, but did not contain exons encoding the transmembrane and cytoplasmic domains, and 3'-untranslated region. The missing portion of the MAdCAM-1 gene was located on clone HEBBC23592, isolated by screening plaques from a FIX II genomic library with a 1.3 kb MAdCAM-1 cDNA probe. Southern blot, PCR, and DNA sequence analysis demonstrated that clone HEBBC23592 contained at least exons 3 to 5 of the MAdCAM-1 gene.

DNA sequencing revealed that the coding portion of the MAdCAM-1 gene is contained within 5 exons, with the sequences being identical to the MAdCAM-1 cDNA sequence. All intron-exon splice junction sequences are in agreement with the GT/AG rule for splicing. The introns are all type I, where interruption occurs after the first nucleotide of a codon. The first exon (52 bp) encodes the signal peptide and 5'-untranslated sequence; exons 2 and 3 encode the N-terminal Ig domains; exon 4 encodes the mucin domain; and exon 5 encodes the transmembrane and cytoplasmic domains, and the 3' untranslated region.

Comparison of the human and mouse MAdCAM-1 genes

Alignment of the human and mouse MAdCAM-1 gene sequences revealed that three of four intron-exon junctions separating the signal peptide and Ig domain sequences were conserved in position. The MAdCAM-1 mucin-like domains are not conserved between species, and the exon-intron splice sites separating the mucin and transmembrane domain sequences are also not conserved. In humans, the splice site is nine amino acids N-terminal to the boundary of the extracellular and transmembrane domains, whereas in mouse it is three amino acids N-terminal to that boundary.

A splice variant of human MAdCAM-1 lacks exon 4 encoding the mucin domain

Splice variants of human MAdCAM-1, where the variant forms lack all or part of the second Ig domain, and all or part of the major mucin domain, are described above in Example 5. Comparison with the MAdCAM-1 genomic sequence confirms that all four splice variants were derived by internal splicing of exons, unlike the single splice variant identified for mouse MAdCAM-1 which is formed by splicing out exon 4, which encodes the mucin/IgA-like Ig domain. Further splice variants of 250 (minor), 350 (major), and 500 (minor) bp in size, compared to a full-length PCR product of 700 bp, were amplified from human fetal brain. Shotgun subcloning and sequencing revealed an equivalent of the mouse exon 4 splice variant, encoded by 340 bp of DNA. Comparison with the genomic sequence reveals that this new MAdCAM-1 variant is created by alternative splicing and deletion of exon 4, which encodes the entire mucin-like domain.

Analysis of the 5'-flanking region of the MAdCAM-1 gene

A 700 bp 5'-flanking region of the MAdCAM-1 gene was sequenced, revealing several potential transcriptional regulatory elements. These include two tandem NF-kB binding sites at positions -98 and -110 with respect to the translational start codon; thirteen SP-1 sites at -66, -141, -157, -164, -177, -189, -308, -322, -338, -590, -647, -664, -678 ; nine AP-2 sites at -66, -157, 204, -325, -544, -549, -694, -591, -204; PEA3 (ets family) sites at -115, -212; an NF-E1 site at -522; Adh1 (ETF) sites at -95, -187; a GC box at -176; a MyoD site at -582; an E2A site at -85; an ENKCRE (SEQ ID NO:43) site at -496; and an IRS site at -354. Only the tandem NF-kB sites, the SP-1 site at -590, and a potential TATA box (TATTTAA; at position -38) (SEQ ID NO:44) identified in the mouse promoter are conserved in position (Fig. 13). Despite this, the 367 bp promoter region immediately flanking the MAdCAM-1 gene is highly conserved (79 %) with the corresponding region of the mouse promoter (Fig. 13).

The pGL-2/B-718+ and pGL-2/B-718- constructs which contain a 700 bp fragment of the MAdCAM-1 gene 5'-flanking sequence (nucleotide positions -718 to +20 relative to the translational start) fused to the luciferase reporter gene (Figs.14A) were used in transient transfection assays to test for promoter activity. Promoter activity was tested in PMA-treated and untreated HMEC cells, a human dermal endothelial cell line which consitutively produces MAdCAM-1 RNA (Fig.14B). The reporter construct directed a low but consistent level of luciferase activity in unactivated cells as compared to the pGL-2/B basic control vector, and the control pGL-2/-718- vector containing the promoter in the incorrect orientation. The activity of the pGL-2/B-718+ vector was doubled

following cell stimulation with PMA, in comparison to the pGL-2/-718 vector control (Fig.14C).

Chromosomal assignment of the human MAdCAM-1 gene

Genomic DNAs from a panel of 24 human-rodent somatic cell hybrids, the majority of which were monochromosomal, were analyzed by PCR using PCR primers directed to the MAdCAM-1 sequence. The expected 386 bp PCR fragment was specifically amplified from human DNA, but not from mouse or hamster DNAs, and was specifically obtained from a hybrid cell line (GM10612) containing only human chromosome 19.

The MAdCAM-1 gene was regionally localized to chromosome 19 by *in situ* hybridisation of metaphase chromosomal spreads with the 1.3 kb cDNA insert of MAdCAM-1 cDNA clone HEBBC23X (see Example 5).Approximately thirteen spreads were analyzed by eye, most of which had a doublet signal characteristic of genuine hybridization on at least one chromosome 19. Doublet signals were not detected on any other chromosome. Detailed analysis of 12 individual chromosomes, using fluorescence banding cimbined with high resolution image analysis, indicated that the MAdCAM-1 gene is positioned within band 19p13.3.

Discussion

The genomic organization of the MAdCAM-1 gene correlates well with the subdomain structure of the encoded protein. The 5'-untranslated region and signal peptide are encoded by exon 1, the two N-terminal Ig domains and mucin-like domain are encoded by exons 2, 3, and 4, respectively, and the

transmembrane and cytoplasmic domains and 3'-untranslated region are combined together on exon 5. Several features of MAdCAM-1 have been conserved between humans and mice, including the structure of the two Ig ligand-binding domains, yet the 3'-region is quite divergent. Comparison of the human gene sequence with the mouse homologue revealed that differences in organization of the 3'-region are not simply due to alternative splicing, but are inherent in the genomic DNA. Thus the human MAdCAM-1 gene contains no sequence equivalent to the third IgA-homologous domain of mouse MAdCAM-1 adjoining the 3'-end of the mucin domain. It is possible that a third Ig domain exists as a separate exon in the large intron separating exons 4 and 5, but given all the available evidence, and in particular sequence analysis of MAdCAM-1 splice variants from RT-PCR analysis, this seems unlikely. Despite this major difference in gross structure other regions of human and mouse MAdCAM-1 are highly conserved, including the positions of four of the five intron-exon splice junctions, highlighting the close evolutionary relationship between the molecules.

Four splice variants were identified by RT-PCT that lacked all or part of the second Ig domain, and all or part of the major mucin domain. Comparison with the genomic sequence reveals that all the variants arose by internal splicing of exons. Intra-exonic splicing of MAdCAM-1 is further substantiated by the fact that our original MAdCAM-1 cDNA clone HEBBC23X has only six major mucin repeats, whereas a human MAdCAM-1 clone isolated from a mesenteric lymph node library contained eight such repeats. A MAdCAM-1 variant containing just six repeats has also been independently isolated by RT-PCR. It was therefore of interest to determine that the total possible number of repeats in the major mucin domain, contained within exon 4 of the MAdCAM-1 gene, is in fact eight. The regulation of mucin adhesion by alternative splicing is well established, and MAdCAM-1 appears to be no exception. The human MAdCAM-1 variant

created by the splicing out of exon 4 encoding the mucin domain (described above in Example 5) is a counterpart to the splice variant identified in mouse MAdCAM-1 which lacks exon 4, encoding the mucin and third IgA-like domain. Despite the prominence of the splice variants identified by PCR, they are not abundant in Northerns. Nevertheless it will be important to study the topographical and tissue distribution of the various MAdCAM-1 splice variants, given that absence or truncations of the mucin domain will affect the ability of MAdCAM-1 to facilitate lymphocyte tethering under flow to L-selectin.

Sequence analysis of the 5'-region of the human MAdCAM-1 gene revealed close similarity to the mouse MAdCAM-1 gene promoter. The two tandem NF-kB sites located 100 bp upstream of the start site of transcription in the mouse promoter are conserved in position. Transfection assays in the murine endothelial cell line bEnd.3, carried out with promoter mutants of the mouse MAdCAM-1 gene, revealed that occupancy of both NF-kB sites is essential for the promoter to drive expression in response to TNF-α. The 5' NF-kB site is totally conserved in sequence with the mouse counterpart, whereas the 3' site is only slightly divergent. NF-kB is also involved in the increased expression of VCAM-1 and ICAM-1 by LPS, TNF-α and IL-1β. In contrast, binding sites for TGF-β-inducible transcription factors (NF1 and AP1), previously identified in the mouse promoter, were not present. Multiple AP-2 sites in addition to the NF-kB sites may be responsible for the increased activity of the promoter in response to PMA. The presence of a MyoD site (CACCTG) (SEQ ID NO:45), which is found within the muscle creatine kinase enhancer, is interesting, given that the related VCAM-1 is expressed on myoblasts and myotubes in culture and *in vivo* at sites of secondary myogenesis.

FISH and PCR analysis of a panel of human-rodent somatic cell hybrids was used to localize the MAdCAM-1 gene to chromosome 19, band p13.3. It is notable that the human ICAM-1 and ICAM-3 genes are located in close proximity (19p13.2-p13.3), raising the possibility that the MAdCAM-1, ICAM-1 and ICAM-3 genes are clustered together on the short arm of chromosome 19. This region is homologous to a region on mouse chromosome 10, and it is interesting therefore that the mouse MAdCAM-1 gene is located on chromosome 10. Yet another member of the immunoglobulin superfamily which is ubiquitously expressed in various tissues, termed basigin, also maps to this same region. In contrast VCAM-1 and ICAM-2 are located on chromosomes 1 and 17, respectively. Given that the MAdCAM-1 mucin-like domain is decorated with carbohydrate moieties recognized by L-selectin, it is interesting to note that a cluster of three (FUT6-FUT3-FUT5) of five cloned human fucosyltransferase genes responsible for the synthesis of sialyl Lewis x and a, and related fucosylated antigens recognised by selectins, is located on 19p13.3. In terms of cancer, band 19p13.3 is frequently involved in structural anomalies of chromosome 19, associated with ovarian cancer, leukemia, and multiple myeloma. Genes at 19p13.3 which have so far been shown to be involved include the insulin receptor, E2A transcription factor, and MLLT1 genes.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: HUMAN GENOME SCIENCES, INC.
9410 KEY WEST AVENUE
ROCKVILLE, MD 20850
UNITED STATES OF AMERICA

THE UNIVERSITY OF AUCKLAND
85 PARK ROAD, GRAFTON
AUCKLAND, NEW ZEALAND

APPLICANTS/INVENTORS: NI, JIAN
GREENE, JOHN M.
KRISSANSEN, GEOFFREY W
LEUNG, EUPHEMIA YEE FUN
RUBEN, STEVEN M.

(ii) TITLE OF INVENTION: HUMAN MUCOSAL ADDRESSIN CELL ADHESION MOLECULE-1 (MAdCAM-1) AND SPLICE VARIANTS THEREOF

(iii) NUMBER OF SEQUENCES: 59

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: STERNE, KESSLER, GOLDSTEIN & FOX P.L.L.C.
(B) STREET: 1100 NEW YORK AVENUE, N.W. SUITE 600
(C) CITY: WASHINGTON
(D) STATE: D.C.
(E) COUNTRY: US
(F) ZIP: 20005

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.30

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: TBA
(B) FILING DATE: HEREWITH
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: GOLDSTEIN, JORGE A.
(B) REGISTRATION NUMBER: 29,021
(C) REFERENCE/DOCKET NUMBER: 1488.057PC00/JAG/EKS/LLK

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 202-371-2600
(B) TELEFAX: 202-371-2540

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1536 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 1..1146

(ix) FEATURE:
(A) NAME/KEY: mat_peptide
(B) LOCATION: 52..1146

(ix) FEATURE:
(A) NAME/KEY: sig_peptide
(B) LOCATION: 1..49

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
ATG GAT TTC GGA CTG GCC CTC CTG CTG GCG GGG CTT CTG GGG CTC CTC          48
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                 -5

CTC GGC CAG TCC CTC CAG GTG AAG CCC CTG CAG GTG GAG CCC CCG GAG          96
Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

CCG GTG GTG GCC GTG GCC TTG GGC GCC TCG CGC CAG CTC ACC TGC CGC         144
Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
              20                  25                  30

CTG GCC TGC GCG GAC CGC GGG GCC TCG GTG CAG TGG CGG GGC CTG GAC         192
Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
          35                  40                  45

ACC AGC CTG GGC GCG GTG CAG TCG GAC ACG GGC CGC AGC GTC CTC ACC         240
Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
      50                  55                  60

GTG CGC AAC GCC TCG CTG TCG GCG GCC GGG ACC CGC GTG TGC GTG GGC         288
Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
  65                  70                  75

TCC TGC GGG GGC CGC ACC TTC CAG CAC ACC GTG CAG CTC CTT GTG TAC         336
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
 80                  85                  90                  95

GCC TTC CCG GAC CAG CTG ACC GTC TCC CCA GCA GCC CTG GTG CCT GGT         384
```

```
Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
                100                 105                 110

GAC CCG GAG GTG GCC TGT ACG GCC CAC AAA GTC ACG CCC GTG GAC CCC      432
Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
            115                 120                 125

AAC GCG CTC TCC TTC TCC CTG CTC GTC GGG GGC CAG GAA CTG GAG GGG      480
Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
        130                 135                 140

GCG CAA GCC CTG GGC CCG GAG GTG CAG GAG GAG GAG GAG GAG CCC CAG      528
Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Glu Glu Glu Glu Pro Gln
    145                 150                 155

GGG GAC GAG GAC GTG CTG TTC AGG GTG ACA GAG CGC TGG CGG CTG CCG      576
Gly Asp Glu Asp Val Leu Phe Arg Val Thr Glu Arg Trp Arg Leu Pro
160                 165                 170                 175

CCC CTG GGG ACC CCT GTC CCG CCC GCC CTC TAC TGC CAG GCC ACG ATG      624
Pro Leu Gly Thr Pro Val Pro Pro Ala Leu Tyr Cys Gln Ala Thr Met
                180                 185                 190

AGG CTG CCT GGC TTG GAG CTC AGC CAC CGC CAG GCC ATC CCC GTC CTG      672
Arg Leu Pro Gly Leu Glu Leu Ser His Arg Gln Ala Ile Pro Val Leu
            195                 200                 205

CAC AGC CCG ACC TCC CCG GAG CCT CCC GAC ACC ACC TCC CCG GAG TCT      720
His Ser Pro Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser Pro Glu Ser
        210                 215                 220

CCC GAC ACC ACC TCC CCG GAG TCT CCC GAC ACC ACC TCC CCG GAG CCT      768
Pro Asp Thr Thr Ser Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro
    225                 230                 235

CCC GAC ACC ACC TCC CCG GAG CCT CCC GAC AAG ACC TCC CCG GAG CCC      816
Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro
240                 245                 250                 255

GCC CCC CAG CAG GGC TCC ACA CAC ACC CCC AGG AGC CCA GGC TCC ACC      864
Ala Pro Gln Gln Gly Ser Thr His Thr Pro Arg Ser Pro Gly Ser Thr
                260                 265                 270

AGG ACT CGC CGC CCT GAG ATC TCC CAG GCT GGG CCC ACG CAG GGA GAA      912
Arg Thr Arg Arg Pro Glu Ile Ser Gln Ala Gly Pro Thr Gln Gly Glu
            275                 280                 285

GTG ATC CCA ACA GGC TCG TCC AAA CCT GCG GGT GAC CAG CTG CCC GCG      960
Val Ile Pro Thr Gly Ser Ser Lys Pro Ala Gly Asp Gln Leu Pro Ala
        290                 295                 300

GCT CTG TGG ACC AGC AGT GCG GTG CTG GGA CTG CTG CTC CTG GCC TTG     1008
Ala Leu Trp Thr Ser Ser Ala Val Leu Gly Leu Leu Leu Leu Ala Leu
    305                 310                 315
```

```
CCC ACG TAT CAC CTC TGG AAA CGC TGC CGG CAC CTG GCT GAG GAC GAC      1056
Pro Thr Tyr His Leu Trp Lys Arg Cys Arg His Leu Ala Glu Asp Asp
320                 325                 330                 335

ACC CAC CCA CCA GCT TCT CTG AGG CTT CTG CCC CAG GTG TCG GCC TGG      1104
Thr His Pro Pro Ala Ser Leu Arg Leu Leu Pro Gln Val Ser Ala Trp
                340                 345                 350

GCT GGG TTA AGG GGG ACC GGC CAG GTC GGG ATC AGC CCC TCC              1146
Ala Gly Leu Arg Gly Thr Gly Gln Val Gly Ile Ser Pro Ser
            355                 360                 365

TGAGTGGCCA GCCTTTCCCC CTGTGAAAGC AAAATAGCTT GGACCCCTTC AAGTTGAGAA    1206

CTGGTCAGGG CAAACCTGCC TCCCATTCTA CTCAAAGTCA TCCCTCTGTT CACAGAGATG    1266

GATGCATGTT CTGATTGCCT CTTTGGAGAA GCTCATCAGA AACTCAAAAG AAGGCCACTG    1326

TTTGTCTCAC CTACCCATGA CCTGAAGCCC CTCCCTGAGT GGTCCCCACC TTTCTGGACG    1386

GAACCACGTA CTTTTTACAT ACATTGATTC ATGTCTCACG TCTCCCTAAA AATGCGTAAG    1446

ACCAAGCTGT GCCCTGACCA CCCTGGGCCC CTGTCGTCAG GACCTCCTGA GGCTTTGGCA    1506

AATAAACCTC CTAAAATGAT AAAAAAAAAA                                     1536
```

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 382 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                 -5

Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
                20                  25                  30

Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
            35                  40                  45

Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
        50                  55                  60

Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
    65                  70                  75
```

```
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
 80                  85                  90                  95

Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
                100                 105                 110

Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
            115                 120                 125

Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
        130                 135                 140

Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Glu Glu Glu Glu Pro Gln
    145                 150                 155

Gly Asp Glu Asp Val Leu Phe Arg Val Thr Glu Arg Trp Arg Leu Pro
160                 165                 170                 175

Pro Leu Gly Thr Pro Val Pro Pro Ala Leu Tyr Cys Gln Ala Thr Met
                180                 185                 190

Arg Leu Pro Gly Leu Glu Leu Ser His Arg Gln Ala Ile Pro Val Leu
            195                 200                 205

His Ser Pro Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser Pro Glu Ser
        210                 215                 220

Pro Asp Thr Thr Ser Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro
    225                 230                 235

Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro
240                 245                 250                 255

Ala Pro Gln Gln Gly Ser Thr His Thr Pro Arg Ser Pro Gly Ser Thr
                260                 265                 270

Arg Thr Arg Arg Pro Glu Ile Ser Gln Ala Gly Pro Thr Gln Gly Glu
            275                 280                 285

Val Ile Pro Thr Gly Ser Ser Lys Pro Ala Gly Asp Gln Leu Pro Ala
        290                 295                 300

Ala Leu Trp Thr Ser Ser Ala Val Leu Gly Leu Leu Leu Leu Ala Leu
    305                 310                 315

Pro Thr Tyr His Leu Trp Lys Arg Cys Arg His Leu Ala Glu Asp Asp
320                 325                 330                 335

Thr His Pro Pro Ala Ser Leu Arg Leu Leu Pro Gln Val Ser Ala Trp
                340                 345                 350

Ala Gly Leu Arg Gly Thr Gly Gln Val Gly Ile Ser Pro Ser
            355                 360                 365
```

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1488 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 1..1098

(ix) FEATURE:
(A) NAME/KEY: mat_peptide
(B) LOCATION: 52..1098

(ix) FEATURE:
(A) NAME/KEY: sig_peptide
(B) LOCATION: 1..49

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

```
ATG GAT TTC GGA CTG GCC CTC CTG CTG GCG GGG CTT CTG GGG CTC CTC        48
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                 -5

CTC GGC CAG TCC CTC CAG GTG AAG CCC CTG CAG GTG GAG CCC CCG GAG        96
Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

CCG GTG GTG GCC GTG GCC TTG GGC GCC TCG CGC CAG CTC ACC TGC CGC       144
Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
              20                  25                  30

CTG GCC TGC GCG GAC CGC GGG GCC TCG GTG CAG TGG CGG GGC CTG GAC       192
Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
          35                  40                  45

ACC AGC CTG GGC GCG GTG CAG TCG GAC ACG GGC CGC AGC GTC CTC ACC       240
Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
      50                  55                  60

GTG CGC AAC GCC TCG CTG TCG GCG GCC GGG ACC CGC GTG TGC GTG GGC       288
Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
  65                  70                  75

TCC TGC GGG GGC CGC ACC TTC CAG CAC ACC GTG CAG CTC CTT GTG TAC       336
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
80                  85                  90                  95

GCC TTC CCG GAC CAG CTG ACC GTC TCC CCA GCA GCC CTG GTG CCT GGT       384
Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
              100                 105                 110
```

```
GAC CCG GAG GTG GCC TGT ACG GCC CAC AAA GTC ACG CCC GTG GAC CCC        432
Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
            115                 120                 125

AAC GCG CTC TCC TTC TCC CTG CTC GTC GGG GGC CAG GAA CTG GAG GGG        480
Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
        130                 135                 140

GCG CAA GCC CTG GGC CCG GAG GTG CAG GAG GAG GAG GAG GAG CCC CAG        528
Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Glu Glu Glu Glu Pro Gln
    145                 150                 155

GGG GAC GAG GAC GTG CTG TTC AGG GTG ACA GAG CGC TGG CGG CTG CCG        576
Gly Asp Glu Asp Val Leu Phe Arg Val Thr Glu Arg Trp Arg Leu Pro
160                 165                 170                 175

CCC CTG GGG ACC CCT GTC CCG CCC GCC CTC TAC TGC CAG GCC ACG ATG        624
Pro Leu Gly Thr Pro Val Pro Pro Ala Leu Tyr Cys Gln Ala Thr Met
                180                 185                 190

AGG CTG CCT GGC TTG GAG CTC AGC CAC CGC CAG GCC ATC CCC GTC CTG        672
Arg Leu Pro Gly Leu Glu Leu Ser His Arg Gln Ala Ile Pro Val Leu
            195                 200                 205

CAC AGC CCG ACC TCC CCG GAG TCT CCC GAC ACC ACC TCC CCG GAG CCT        720
His Ser Pro Thr Ser Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro
        210                 215                 220

CCC GAC ACC ACC TCC CCG GAG CCT CCC GAC AAG ACC TCC CCG GAG CCC        768
Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro
    225                 230                 235

GCC CCC CAG CAG GGC TCC ACA CAC ACC CCC AGG AGC CCA GGC TCC ACC        816
Ala Pro Gln Gln Gly Ser Thr His Thr Pro Arg Ser Pro Gly Ser Thr
240                 245                 250                 255

AGG ACT CGC CGC CCT GAG ATC TCC CAG GCT GGG CCC ACG CAG GGA GAA        864
Arg Thr Arg Arg Pro Glu Ile Ser Gln Ala Gly Pro Thr Gln Gly Glu
                260                 265                 270

GTG ATC CCA ACA GGC TCG TCC AAA CCT GCG GGT GAC CAG CTG CCC GCG        912
Val Ile Pro Thr Gly Ser Ser Lys Pro Ala Gly Asp Gln Leu Pro Ala
            275                 280                 285

GCT CTG TGG ACC AGC AGT GCG GTG CTG GGA CTG CTG CTC CTG GCC TTG        960
Ala Leu Trp Thr Ser Ser Ala Val Leu Gly Leu Leu Leu Leu Ala Leu
        290                 295                 300

CCC ACG TAT CAC CTC TGG AAA CGC TGC CGG CAC CTG GCT GAG GAC GAC       1008
Pro Thr Tyr His Leu Trp Lys Arg Cys Arg His Leu Ala Glu Asp Asp
    305                 310                 315

ACC CAC CCA CCA GCT TCT CTG AGG CTT CTG CCC CAG GTG TCG GCC TGG       1056
Thr His Pro Pro Ala Ser Leu Arg Leu Leu Pro Gln Val Ser Ala Trp
320                 325                 330                 335
```

```
GCT GGG TTA AGG GGG ACC GGC CAG GTC GGG ATC AGC CCC TCC                1098
Ala Gly Leu Arg Gly Thr Gly Gln Val Gly Ile Ser Pro Ser
                340                 345

TGAGTGGCCA GCCTTTCCCC CTGTGAAAGC AAAATAGCTT GGACCCCTTC AAGTTGAGAA     1158

CTGGTCAGGG CAAACCTGCC TCCCATTCTA CTCAAAGTCA TCCCTCTGTT CACAGAGATG     1218

GATGCATGTT CTGATTGCCT CTTTGGAGAA GCTCATCAGA AACTCAAAAG AAGGCCACTG     1278

TTTGTCTCAC CTACCCATGA CCTGAAGCCC CTCCCTGAGT GGTCCCCACC TTTCTGGACG     1338

GAACCACGTA CTTTTTACAT ACATTGATTC ATGTCTCACG TCTCCCTAAA AATGCGTAAG     1398

ACCAAGCTGT GCCCTGACCA CCCTGGGCCC CTGTCGTCAG GACCTCCTGA GGCTTTGGCA     1458

AATAAACCTC CTAAAATGAT AAAAAAAAAA                                      1488
```

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 366 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                  -5

Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
      1               5                  10                  15

Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
                20                  25                  30

Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
            35                  40                  45

Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
        50                  55                  60

Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
    65                  70                  75

Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
 80                  85                  90                  95

Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
                100                 105                 110

Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
```

115 120 125

Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
130 135 140

Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Glu Glu Glu Glu Pro Gln
145 150 155

Gly Asp Glu Asp Val Leu Phe Arg Val Thr Glu Arg Trp Arg Leu Pro
160 165 170 175

Pro Leu Gly Thr Pro Val Pro Pro Ala Leu Tyr Cys Gln Ala Thr Met
180 185 190

Arg Leu Pro Gly Leu Glu Leu Ser His Arg Gln Ala Ile Pro Val Leu
195 200 205

His Ser Pro Thr Ser Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro
210 215 220

Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro
225 230 235

Ala Pro Gln Gln Gly Ser Thr His Thr Pro Arg Ser Pro Gly Ser Thr
240 245 250 255

Arg Thr Arg Arg Pro Glu Ile Ser Gln Ala Gly Pro Thr Gln Gly Glu
260 265 270

Val Ile Pro Thr Gly Ser Ser Lys Pro Ala Gly Asp Gln Leu Pro Ala
275 280 285

Ala Leu Trp Thr Ser Ser Ala Val Leu Gly Leu Leu Leu Leu Ala Leu
290 295 300

Pro Thr Tyr His Leu Trp Lys Arg Cys Arg His Leu Ala Glu Asp Asp
305 310 315

Thr His Pro Pro Ala Ser Leu Arg Leu Leu Pro Gln Val Ser Ala Trp
320 325 330 335

Ala Gly Leu Arg Gly Thr Gly Gln Val Gly Ile Ser Pro Ser
340 345

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1179 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 1..789

(ix) FEATURE:
(A) NAME/KEY: mat_peptide
(B) LOCATION: 52..789

(ix) FEATURE:
(A) NAME/KEY: sig_peptide
(B) LOCATION: 1..49

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

```
ATG GAT TTC GGA CTG GCC CTC CTG CTG GCG GGG CTT CTG GGG CTC CTC        48
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                 -5

CTC GGC CAG TCC CTC CAG GTG AAG CCC CTG CAG GTG GAG CCC CCG GAG        96
Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

CCG GTG GTG GCC GTG GCC TTG GGC GCC TCG CGC CAG CTC ACC TGC CGC       144
Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
                20                  25                  30

CTG GCC TGC GCG GAC CGC GGG GCC TCG GTG CAG TGG CGG GGC CTG GAC       192
Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
            35                  40                  45

ACC AGC CTG GGC GCG GTG CAG TCG GAC ACG GGC CGC AGC GTC CTC ACC       240
Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
        50                  55                  60

GTG CGC AAC GCC TCG CTG TCG GCG GCC GGG ACC CGC GTG TGC GTG GGC       288
Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
    65                  70                  75

TCC TGC GGG GGC CGC ACC TTC CAG CAC ACC GTG CAG CTC CTT GTG TAC       336
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
80                  85                  90                  95

GCC TTC CCG GAC CAG CTG ACC GTC TCC CCA GCA GCC CTG GTG CCT GGT       384
Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
                100                 105                 110

GAC CCG GAG GTG GCC TGT ACG GCC CAC AAA GTC ACG CCC GTG GAC CCC       432
Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
            115                 120                 125

AAC GCG CTC TCC TTC TCC CTG CTC GTC GGG GGC CAG CAG GGC TCC ACA       480
Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Gln Gly Ser Thr
        130                 135                 140
```

```
CAC ACC CCC AGG AGC CCA GGC TCC ACC AGG ACT CGC CGC CCT GAG ATC       528
His Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro Glu Ile
    145                 150                 155

TCC CAG GCT GGG CCC ACG CAG GGA GAA GTG ATC CCA ACA GGC TCG TCC       576
Ser Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly Ser Ser
160                 165                 170                 175

AAA CCT GCG GGT GAC CAG CTG CCC GCG GCT CTG TGG ACC AGC AGT GCG       624
Lys Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser Ser Ala
                180                 185                 190

GTG CTG GGA CTG CTG CTC CTG GCC TTG CCC ACG TAT CAC CTC TGG AAA       672
Val Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu Trp Lys
            195                 200                 205

CGC TGC CGG CAC CTG GCT GAG GAC GAC ACC CAC CCA CCA GCT TCT CTG       720
Arg Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala Ser Leu
        210                 215                 220

AGG CTT CTG CCC CAG GTG TCG GCC TGG GCT GGG TTA AGG GGG ACC GGC       768
Arg Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly Thr Gly
    225                 230                 235

CAG GTC GGG ATC AGC CCC TCC TGAGTGGCCA GCCTTTCCCC CTGTGAAAGC          819
Gln Val Gly Ile Ser Pro Ser
240                 245

AAAATAGCTT GGACCCCTTC AAGTTGAGAA CTGGTCAGGG CAAACCTGCC TCCCATTCTA     879

CTCAAAGTCA TCCCTCTGTT CACAGAGATG GATGCATGTT CTGATTGCCT CTTTGGAGAA     939

GCTCATCAGA AACTCAAAAG AAGGCCACTG TTTGTCTCAC CTACCCATGA CCTGAAGCCC     999

CTCCCTGAGT GGTCCCCACC TTTCTGGACG GAACCACGTA CTTTTTACAT ACATTGATTC    1059

ATGTCTCACG TCTCCCTAAA AATGCGTAAG ACCAAGCTGT GCCCTGACCA CCCTGGGCCC    1119

CTGTCGTCAG GACCTCCTGA GGCTTTGGCA AATAAACCTC CTAAAATGAT AAAAAAAAAA    1179
```

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 263 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

```
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17      -15                 -10                  -5
```

```
Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
 1               5                  10                  15

Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
            20                  25                  30

Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
        35                  40                  45

Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
    50                  55                  60

Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
65                  70                  75

Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
80                  85                  90                  95

Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
            100                 105                 110

Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
        115                 120                 125

Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Gln Gly Ser Thr
    130                 135                 140

His Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro Glu Ile
145                 150                 155

Ser Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly Ser Ser
160                 165                 170                 175

Lys Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser Ser Ala
            180                 185                 190

Val Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu Trp Lys
        195                 200                 205

Arg Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala Ser Leu
    210                 215                 220

Arg Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly Thr Gly
225                 230                 235

Gln Val Gly Ile Ser Pro Ser
240                 245
```

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1320 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 1..930

(ix) FEATURE:
(A) NAME/KEY: mat_peptide
(B) LOCATION: 52..930

(ix) FEATURE:
(A) NAME/KEY: sig_peptide
(B) LOCATION: 1..49

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
ATG GAT TTC GGA CTG GCC CTC CTG CTG GCG GGG CTT CTG GGG CTC CTC        48
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                 -5

CTC GGC CAG TCC CTC CAG GTG AAG CCC CTG CAG GTG GAG CCC CCG GAG        96
Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

CCG GTG GTG GCC GTG GCC TTG GGC GCC TCG CGC CAG CTC ACC TGC CGC       144
Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
              20                  25                  30

CTG GCC TGC GCG GAC CGC GGG GCC TCG GTG CAG TGG CGG GGC CTG GAC       192
Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
          35                  40                  45

ACC AGC CTG GGC GCG GTG CAG TCG GAC ACG GGC CGC AGC GTC CTC ACC       240
Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
      50                  55                  60

GTG CGC AAC GCC TCG CTG TCG GCG GCC GGG ACC CGC GTG TGC GTG GGC       288
Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
  65                  70                  75

TCC TGC GGG GGC CGC ACC TTC CAG CAC ACC GTG CAG CTC CTT GTG TAC       336
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
80                  85                  90                  95

GCC TTC CCG GAC CAG CTG ACC GTC TCC CCA GCA GCC CTG GTG CCT GGT       384
Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
              100                 105                 110

GAC CCG GAG GTG GCC TGT ACG GCC CAC AAA GTC ACG CCC GTG GAC CCC       432
Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
          115                 120                 125

AAC GCG CTC TCC TTC TCC CTG CTC GTC GGG GGC CAG GAA CTG GAG GGG       480
```

```
Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
        130                 135                 140

GCG CAA GCC CTG GGC CCG GAG TCT CCC GAC ACC ACC TCC CCG GAG TCT      528
Ala Gln Ala Leu Gly Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Ser
    145                 150                 155

CCC GAC ACC ACC TCC CCG GAG CCT CCC GAC ACC ACC TCC CCG GAG CCT      576
Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser Pro Glu Pro
160                 165                 170                 175

CCC GAC AAG ACC TCC CCG GAG CCC GCC CCC CAG CAG GGC TCC ACA CAC      624
Pro Asp Lys Thr Ser Pro Glu Pro Ala Pro Gln Gln Gly Ser Thr His
                180                 185                 190

ACC CCC AGG AGC CCA GGC TCC ACC AGG ACT CGC CGC CCT GAG ATC TCC      672
Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro Glu Ile Ser
            195                 200                 205

CAG GCT GGG CCC ACG CAG GGA GAA GTG ATC CCA ACA GGC TCG TCC AAA      720
Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly Ser Ser Lys
        210                 215                 220

CCT GCG GGT GAC CAG CTG CCC GCG GCT CTG TGG ACC AGC AGT GCG GTG      768
Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser Ser Ala Val
    225                 230                 235

CTG GGA CTG CTG CTC CTG GCC TTG CCC ACG TAT CAC CTC TGG AAA CGC      816
Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu Trp Lys Arg
240                 245                 250                 255

TGC CGG CAC CTG GCT GAG GAC GAC ACC CAC CCA CCA GCT TCT CTG AGG      864
Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala Ser Leu Arg
                260                 265                 270

CTT CTG CCC CAG GTG TCG GCC TGG GCT GGG TTA AGG GGG ACC GGC CAG      912
Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly Thr Gly Gln
            275                 280                 285

GTC GGG ATC AGC CCC TCC TGAGTGGCCA GCCTTTCCCC CTGTGAAAGC             960
Val Gly Ile Ser Pro Ser
        290

AAAATAGCTT GGACCCCTTC AAGTTGAGAA CTGGTCAGGG CAAACCTGCC TCCCATTCTA    1020

CTCAAAGTCA TCCCTCTGTT CACAGAGATG GATGCATGTT CTGATTGCCT CTTTGGAGAA    1080

GCTCATCAGA AACTCAAAAG AAGGCCACTG TTTGTCTCAC CTACCCATGA CCTGAAGCCC    1140

CTCCCTGAGT GGTCCCCACC TTTCTGGACG GAACCACGTA CTTTTTACAT ACATTGATTC    1200

ATGTCTCACG TCTCCCTAAA AATGCGTAAG ACCAAGCTGT GCCCTGACCA CCCTGGGCCC    1260

CTGTCGTCAG GACCTCCTGA GGCTTTGGCA AATAAACCTC CTAAAATGAT AAAAAAAAAA    1320
```

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 310 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17      -15                 -10                  -5

Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
            20                  25                  30

Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
        35                  40                  45

Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
    50                  55                  60

Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
65                  70                  75

Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
80                  85                  90                  95

Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
            100                 105                 110

Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
        115                 120                 125

Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
    130                 135                 140

Ala Gln Ala Leu Gly Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Ser
145                 150                 155

Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser Pro Glu Pro
160                 165                 170                 175

Pro Asp Lys Thr Ser Pro Glu Pro Ala Pro Gln Gln Gly Ser Thr His
            180                 185                 190

Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro Glu Ile Ser
        195                 200                 205

Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly Ser Ser Lys
```

```
       210                 215                 220

Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser Ser Ala Val
    225                 230                 235

Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu Trp Lys Arg
240                 245                 250                 255

Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala Ser Leu Arg
                260                 265                 270

Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly Thr Gly Gln
            275                 280                 285

Val Gly Ile Ser Pro Ser
        290
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1329 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 1..939

(ix) FEATURE:
(A) NAME/KEY: mat_peptide
(B) LOCATION: 52..939

(ix) FEATURE:
(A) NAME/KEY: sig_peptide
(B) LOCATION: 1..49

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

```
ATG GAT TTC GGA CTG GCC CTC CTG CTG GCG GGG CTT CTG GGG CTC CTC          48
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17     -15                 -10                 -5

CTC GGC CAG TCC CTC CAG GTG AAG CCC CTG CAG GTG GAG CCC CCG GAG          96
Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1                 5                  10                  15

CCG GTG GTG GCC GTG GCC TTG GGC GCC TCG CGC CAG CTC ACC TGC CGC         144
Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
                20                  25                  30
```

```
CTG GCC TGC GCG GAC CGC GGG GCC TCG GTG CAG TGG CGG GGC CTG GAC        192
Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
             35                  40                  45

ACC AGC CTG GGC GCG GTG CAG TCG GAC ACG GGC CGC AGC GTC CTC ACC        240
Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
         50                  55                  60

GTG CGC AAC GCC TCG CTG TCG GCG GCC GGG ACC CGC GTG TGC GTG GGC        288
Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
     65                  70                  75

TCC TGC GGG GGC CGC ACC TTC CAG CAC ACC GTG CAG CTC CTT GTG TAC        336
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
 80                  85                  90                  95

GCC TTC CCG GAC CAG CTG ACC GTC TCC CCA GCA GCC CTG GTG CCT GGT        384
Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
                100                 105                 110

GAC CCG GAG GTG GCC TGT ACG GCC CAC AAA GTC ACG CCC GTG GAC CCC        432
Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
            115                 120                 125

AAC GCG CTC TCC TTC TCC CTG CTC GTC GGG GGC CAG GAA CTG GAG GGG        480
Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
        130                 135                 140

GCG CAA GCC CTG GGC CCG GAG GTG CAG GAG TCT CCC GAC ACC ACC TCC        528
Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Ser Pro Asp Thr Thr Ser
    145                 150                 155

CCG GAG TCT CCC GAC ACC ACC TCC CCG GAG CCT CCC GAC ACC ACC TCC        576
Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser
160                 165                 170                 175

CCG GAG CCT CCC GAC AAG ACC TCC CCG GAG CCC GCC CCC CAG CAG GGC        624
Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro Ala Pro Gln Gln Gly
                180                 185                 190

TCC ACA CAC ACC CCC AGG AGC CCA GGC TCC ACC AGG ACT CGC CGC CCT        672
Ser Thr His Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro
            195                 200                 205

GAG ATC TCC CAG GCT GGG CCC ACG CAG GGA GAA GTG ATC CCA ACA GGC        720
Glu Ile Ser Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly
        210                 215                 220

TCG TCC AAA CCT GCG GGT GAC CAG CTG CCC GCG GCT CTG TGG ACC AGC        768
Ser Ser Lys Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser
    225                 230                 235

AGT GCG GTG CTG GGA CTG CTG CTC CTG GCC TTG CCC ACG TAT CAC CTC        816
Ser Ala Val Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu
240                 245                 250                 255
```

```
TGG AAA CGC TGC CGG CAC CTG GCT GAG GAC GAC ACC CAC CCA CCA GCT      864
Trp Lys Arg Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala
                260                 265                 270

TCT CTG AGG CTT CTG CCC CAG GTG TCG GCC TGG GCT GGG TTA AGG GGG      912
Ser Leu Arg Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly
            275                 280                 285

ACC GGC CAG GTC GGG ATC AGC CCC TCC TGAGTGGCCA GCCTTTCCCC            959
Thr Gly Gln Val Gly Ile Ser Pro Ser
        290                 295

CTGTGAAAGC AAAATAGCTT GGACCCCTTC AAGTTGAGAA CTGGTCAGGG CAAACCTGCC   1019

TCCCATTCTA CTCAAAGTCA TCCCTCTGTT CACAGAGATG GATGCATGTT CTGATTGCCT   1079

CTTTGGAGAA GCTCATCAGA AACTCAAAAG AAGGCCACTG TTTGTCTCAC CTACCCATGA   1139

CCTGAAGCCC CTCCCTGAGT GGTCCCCACC TTTCTGGACG GAACCACGTA CTTTTTACAT   1199

ACATTGATTC ATGTCTCACG TCTCCCTAAA AATGCGTAAG ACCAAGCTGT GCCCTGACCA   1259

CCCTGGGCCC CTGTCGTCAG GACCTCCTGA GGCTTTGGCA AATAAACCTC CTAAAATGAT   1319

AAAAAAAAAA                                                          1329
```

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 313 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

```
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
-17  -15                 -10                  -5

Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
  1               5                  10                  15

Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
                20                  25                  30

Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
            35                  40                  45

Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
        50                  55                  60

Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
    65                  70                  75
```

```
Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
 80                  85                  90                  95

Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
                100                 105                 110

Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
            115                 120                 125

Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
        130                 135                 140

Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Ser Pro Asp Thr Thr Ser
    145                 150                 155

Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser
160                 165                 170                 175

Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro Ala Pro Gln Gln Gly
                180                 185                 190

Ser Thr His Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro
            195                 200                 205

Glu Ile Ser Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly
        210                 215                 220

Ser Ser Lys Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser
    225                 230                 235

Ser Ala Val Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu
240                 245                 250                 255

Trp Lys Arg Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala
                260                 265                 270

Ser Leu Arg Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly
            275                 280                 285

Thr Gly Gln Val Gly Ile Ser Pro Ser
        290                 295
```

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

CGCCCATGGG CCAGTCCCTC CAGGTG 26

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CGCAAGCTTT CAGGGCAGCT GGTCACCCGC 30

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 33 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

CGCGGATCCG CCATCATGGA TTTCGGACTG GCC 33

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

CGCGGTACCT CACTTGAAGG GGTCCAAGC 29

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 33 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

CGCGGATCCG CCATCATGGA TTTCGGACTG GCC 33

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

CGCGGTACCT CAGGGCAGCT GGTCACCCGC 30

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 33 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

CGCGGATCCG CCATCATGGA TTTCGGACTG GCC 33

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

CGCGGTACCT CACTTGAAGG GGTCCAAGC 29

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 33 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

CGCGGATCCG CCATCATGGA TTTCGGACTG GCC 33

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 39 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

CGCTCTAGAT CAAGCGTAGT CTCCGACGTC GTATGGGTA 39

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

CGCGGTACCT CAGGGCAGCT GGTCACCCGC 30

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

CGCTCTCCTT CTCCCTGCTC 20

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

TGGTGGGTGG GTGTCGTCCT C 21

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

CGGCAGCGTT TCCAGAGGTG ATAC 24

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

GGGACTGAGC ATGGATTTCG ACTGGCCCT 29

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 32 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

CGTACAGGCC ACCTCCGGGT CACCAGGCAC CA 32

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 32 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

GCTGGTCCGG GAAGGCGTAC ACAAGGAGCT GC 32

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 7 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

AAATAAA 7

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 7 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: mRNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

RNNAUGR 7

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:
(A) NAME/KEY: Modified-site
(B) LOCATION: 7
(D) OTHER INFORMATION: /note= "CAN BE EITHER PRO OR SER"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

Asp Thr Thr Ser Pro Gly Xaa Pro
1 5

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

Pro Asp Thr Arg Pro Ala Pro Gly Ser Thr Ala Pro Pro Ala His Gly
1 5 10 15

Val Thr Ser Ala
20

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

Thr Thr Thr Pro Asp Val
1 5

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 718 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

CTGCAGCTCC GGAACGGGGG GGGGCTGCTC TCCACCGCCC CTGTGCGGCC GCCCGGGAAA 60

GTGCAGGCGG GCCGGGCGCG GTGGCTCACG CCTGTGATCT CAGCACTTTG GGAGGCCGAG 120

GTGGGCGGAT CACCTGAGGT CGGGAGTTCG AGGCCAGCCT GCCCAACATG GAGAAACCCT 180

GTCTCTACTA AAGATACAAA ATTAGCCAGG CGTGGTGACG CATGCCTGTA ATCCCAGCTA 240

CTGGAGTGGC TGAGGCAGGA GAATCGCTTG AGCCCGGGAG ACAGAGGTTG CGGTGAGCTG 300

AGATCGCACC ATTGCAACTC CAGCCTGGGC AACAAGAGCG AAACTCAGAA AAAAAAGAAA 360

AGAAAGTGCA GGGGACCCGC CGTCGGGGTG GGGGCGGCGC TGCCCAGCCT CTGTCCCACT 420

TCCATGCACT TGACCTCGAC CCTCCGGCCT CCGTCTGCGA TCTTCCCGTG CCTGAATATG 480

AGGCTTGGAA CAGACCCAGA CCTTCCTGCC TGCCCGTCCT GAGTGGCCCC GGGACCCCGC 540

CCCATCTTTG GCCCCCAGCC CCTGCCTTTT TGCCGCCTCC AGGGTCGGGG GTCAGGCCAG 600

GAAAGCCCCT TGGGAAGCCC CCGGGGAGCA GCTGGAGCGG GGTCGCCGGG CGGCGGGAAG 660

GAGTGGGCGC CTCTATTTAA GCGGCTTCCC CGCGGCCTCG GGACAGAGGG GACTGAGC 718

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

ATGGATTTCG GACTGGCCCT CCTGCTGGCG GGGCTTCTGG GGCTCCTCCT CGGTGAGAAG 60

GG 62

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 305 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

GTCGCCGCAG GCCAGTCCCT CCAGGTGAAG CCCCTGCAGG TGGAGCCCCC GGAGCCGGTG 60

GTGGCCGTGG CCTTGGGCGC CTCGCGCCAG CTCACCTGCC GCCTGGCCTG CGCGGACCGC 120

GGGGCCTCGG TGCAGTGGCG GGGCCTGGAC ACCAGCCTGG GCGCGGTGCA GTCGGACACG 180

GGCCGCAGCG TCCTCACCGT GCGCAACGCC TCGCTGTCGG CGGCCGGGAC CCGCGTGTGC 240

GTGGGCTCCT GCGGGGGCCG CACCTTCCAG CACACCGTGC AGCTCCTTGT GTACGGTGAG 300

GCGTC 305

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 350 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

TCCATCACAG CCTTCCCGGA CCAGCTGACC GTCTCCCCAG CAGCCCTGGT GCCTGGTGAC 60

CCGGAGGTGG CCTGTACGGC CCACAAAGTC ACGCCCGTGG ACCCCAACGC GCTCTCCTTC 120

TCCCTGCTCG TCGGGGGCCA GGAACTGGAG GGGGCGCAAG CCCTGGGCCC GGAGGTGCAG 180

GAGGAGGAGG AGGAGCCCCA GGGGGACGAG GACGTGCTGT TCAGGGTGAC AGAGCGCTGG 240

CGGCTGCCGC CCCTGGGGAC CCCTGTCCCG CCCGCCCTCT ACTGCCAGGC CACGATGAGG 300

CTGCCTGGCT TGGAGCTCAG CCACCGCCAG GCCATCCCCG GTGAGTCCGC 350

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 353 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

CTGTTTCCAG TCCTGCACAG CCCGACCTCC CCGGAGCCTC CCGACACCAC CTCCCCGGAG 60

CCTCCCAACA CCACCTCCCC GGAGTCTCCC GACACCACCT CCCCGGAGTC TCCCGACACC 120

ACCTCCCAGG AGCCTCCCGA CACCACCTCC CAGGAGCCTC CCGACACCAC CTCCCAGGAG 180

CCTCCCGACA CCACCTCCCC GGAGCCTCCC GACAAGACCT CCCCGGAGCC CGCCCCCCAG 240

CAGGGCTCCA CACACACCCC CAGGAGCCCA GGCTCCACCA GGACTCGCCG CCCTGAGATC 300

TCCCAGGCTG GGCCCACGCA GGGAGAAGTG ATCCCAACAG GCTGTGAGTT CTG 353

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 608 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CTCTCCCCAG CGTCCAAACC TGCGGGTGAC CAGCTGCCCG CGGCTCTGTG GACCAGCAGT 60

GCGGTGCTGG GACTGCTGCT CCTGGCCTTG CCCACCTATC ACCTCTGGAA ACGCTGCCGG 120

CACCTGGCTG AGGACGACAC CCACCCACCA GCTTCTCTGA GGCTTCTGCC CCAGGTGTCG 180

GCCTGGGCTG GGTTAAGGGG GACCGGCCAG GTCGGGATCA GCCCCTCCTG AGTGGCCAGC 240

CTTTCCCCCT GTGAAAGCAA AATAGCTTGG ACCCCTTCAA GTTGAGAACT GGTCAGGGCA 300

AACCTGCCTC CCATTCTACT CAAAGTCATC CCTCTGTTCA CAGAGATGGA TGCATGTTCT 360

GATTGCCTCT TTGGAGAAGC TCATCAGAAA CTCAAAAGAA GGCCACTGTT TGTCTCACCT 420

ACCCATGACC TGAAGCCCCT CCCTGAGTGG TCCCCACCTT TCTGGACGGA ACCACGTACT 480

TTTTACATAC ATTGATTCAT GTCTCACGTC TCCCTAAAAA TGCGTAAGAC CAAGCTGTGC 540

CCTGACCACC CTGGGCCCCT GTCGTCAGGA CCTCCTGAGG CTTTGGCAAA TAAACCTCCT 600

AAAATGAT 608

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

TGCGGTGCTG GGACTGCTGC TC 22

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

TCAGGGAGGG GCTTCAGGTC A 21

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

GTAATACGAC TCACTATAGG 20

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

AGGGCCAGTC CGAAATCCAT GCTCAGTCCC 30

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

Glu Asn Lys Cys Arg Glu
1 5

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 7 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

TATTTAA 7

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

CACCTG 6

(2) INFORMATION FOR SEQ ID NO:46:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 405 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

Met Glu Ser Ile Leu Ala Leu Leu Leu Ala Leu Ala Leu Val Pro Tyr
1 5 10 15

Gln Leu Ser Arg Gly Gln Ser Phe Gln Val Asn Pro Pro Glu Ser Glu
20 25 30

Val Ala Val Ala Met Gly Thr Ser Leu Gln Ile Thr Cys Ser Met Ser
35 40 45

Cys Asp Glu Gly Val Ala Arg Val His Trp Arg Gly Leu Asp Thr Ser
50 55 60

Leu Gly Ser Val Gln Thr Leu Pro Gly Ser Ser Ile Leu Ser Val Arg
65 70 75 80

Gly Met Leu Ser Asp Thr Gly Thr Pro Val Cys Val Gly Ser Cys Gly
85 90 95

Ser Arg Ser Phe Gln His Ser Val Lys Ile Leu Val Tyr Ala Phe Pro
100 105 110

Asp Gln Leu Val Val Ser Pro Glu Phe Leu Val Pro Gly Gln Asp Gln
115 120 125

Val Val Ser Cys Thr Ala His Asn Ile Trp Pro Ala Asp Pro Asn Ser
130 135 140

Leu Ser Phe Ala Leu Leu Leu Gly Glu Gln Arg Leu Glu Gly Ala Gln
145 150 155 160

Ala Leu Glu Pro Glu Gln Glu Glu Glu Ile Gln Glu Ala Glu Gly Thr
165 170 175

Pro Leu Phe Arg Met Thr Gln Arg Trp Arg Leu Pro Ser Leu Gly Thr
180 185 190

Pro Ala Pro Pro Ala Leu His Cys Gln Val Thr Met Gln Leu Pro Lys
195 200 205

Leu Val Leu Thr His Arg Lys Glu Ile Pro Val Leu Gln Ser Gln Thr

```
    210                 215                 220

Ser Pro Lys Pro Pro Asn Thr Thr Ser Ala Glu Pro Tyr Ile Leu Thr
225                 230                 235                 240

Ser Ser Ser Thr Ala Glu Ala Val Ser Thr Gly Leu Asn Ile Thr Thr
                245                 250                 255

Leu Pro Ser Ala Pro Pro Tyr Pro Lys Leu Ser Pro Arg Thr Leu Ser
            260                 265                 270

Ser Glu Gly Pro Cys Arg Pro Lys Ile His Gln Asp Leu Glu Ala Gly
        275                 280                 285

Trp Glu Leu Leu Cys Glu Ala Ser Cys Gly Pro Gly Val Thr Val Arg
    290                 295                 300

Trp Thr Leu Ala Pro Gly Asp Leu Ala Thr Tyr His Lys Arg Glu Ala
305                 310                 315                 320

Gly Ala Gln Ala Trp Leu Ser Val Leu Pro Pro Gly Pro Met Val Glu
                325                 330                 335

Gly Trp Phe Gln Cys Arg Gln Asp Pro Gly Gly Glu Val Thr Asn Leu
            340                 345                 350

Tyr Val Pro Gly Gln Val Thr Pro Asn Ser Ser Ser Thr Val Val Leu
        355                 360                 365

Trp Ile Gly Ser Leu Val Leu Gly Leu Leu Ala Leu Val Phe Leu Ala
    370                 375                 380

Tyr Arg Leu Trp Lys Cys Tyr Arg Pro Gly Pro Arg Pro Asp Thr Ser
385                 390                 395                 400

Ser Cys Thr His Leu
                405
```

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 406 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

```
Met Asp Phe Gly Leu Ala Leu Leu Leu Ala Gly Leu Leu Gly Leu Leu
1               5                   10                  15
```

Leu Gly Gln Ser Leu Gln Val Lys Pro Leu Gln Val Glu Pro Pro Glu
20 25 30

Pro Val Val Ala Val Ala Leu Gly Ala Ser Arg Gln Leu Thr Cys Arg
35 40 45

Leu Ala Cys Ala Asp Arg Gly Ala Ser Val Gln Trp Arg Gly Leu Asp
50 55 60

Thr Ser Leu Gly Ala Val Gln Ser Asp Thr Gly Arg Ser Val Leu Thr
65 70 75 80

Val Arg Asn Ala Ser Leu Ser Ala Ala Gly Thr Arg Val Cys Val Gly
85 90 95

Ser Cys Gly Gly Arg Thr Phe Gln His Thr Val Gln Leu Leu Val Tyr
100 105 110

Ala Phe Pro Asp Gln Leu Thr Val Ser Pro Ala Ala Leu Val Pro Gly
115 120 125

Asp Pro Glu Val Ala Cys Thr Ala His Lys Val Thr Pro Val Asp Pro
130 135 140

Asn Ala Leu Ser Phe Ser Leu Leu Val Gly Gly Gln Glu Leu Glu Gly
145 150 155 160

Ala Gln Ala Leu Gly Pro Glu Val Gln Glu Glu Glu Glu Glu Pro Gln
165 170 175

Gly Asp Glu Asp Val Leu Phe Arg Val Thr Glu Arg Trp Arg Leu Pro
180 185 190

Pro Leu Gly Thr Pro Val Pro Pro Ala Leu Tyr Cys Gln Ala Thr Met
195 200 205

Arg Leu Pro Gly Leu Glu Leu Ser His Arg Gln Ala Ile Pro Val Leu
210 215 220

His Ser Pro Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser Pro Glu Pro
225 230 235 240

Pro Asn Thr Thr Ser Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Ser
245 250 255

Pro Asp Thr Thr Ser Gln Glu Pro Pro Asp Thr Thr Ser Gln Glu Pro
260 265 270

Pro Asp Thr Thr Ser Gln Glu Pro Pro Asp Thr Thr Ser Pro Glu Pro
275 280 285

Pro Asp Lys Thr Ser Pro Glu Pro Ala Pro Gln Gln Gly Ser Thr His
290 295 300

Thr Pro Arg Ser Pro Gly Ser Thr Arg Thr Arg Arg Pro Glu Ile Ser

305 310 315 320

Gln Ala Gly Pro Thr Gln Gly Glu Val Ile Pro Thr Gly Ser Ser Lys
325 330 335

Pro Ala Gly Asp Gln Leu Pro Ala Ala Leu Trp Thr Ser Ser Ala Val
340 345 350

Leu Gly Leu Leu Leu Leu Ala Leu Pro Thr Tyr His Leu Trp Lys Arg
355 360 365

Cys Arg His Leu Ala Glu Asp Asp Thr His Pro Pro Ala Ser Leu Arg
370 375 380

Leu Leu Pro Gln Val Ser Ala Trp Ala Gly Leu Arg Gly Thr Gly Gln
385 390 395 400

Val Gly Ile Ser Pro Ser
405

(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 408 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

CCCCTCTGCC GCCCTCATGC GGCCACCTGT GGAAGTGAAG GCACAGCTCT AGTCAGCGAG 60

GTGGGCGGGG CAACCTAGGA CTGGCAGATT TCCATGCACT TGACCCACCA TGGTGACCCA 120

CCTCCAGCTT TTAGCTTCAG CCTTCCCGTA CATAGAACCG GGGCCTGGAA CCTTCCCAGA 180

CCTTCCCTCC CCATCTGTAA TGACTGTGTT CCCGGGTCCC TGCCTCACCT CTAGCCTCTG 240

ATTCTCTGCC TCCTACAAAG TGGGGGTCGG GCTGGGAAAG CCCCCTGGGA AAGTCCCACA 300

GAGCCGGCAG AAGGGGGAGG AGAGGCAGGG TCTCAGACAG TAGGAAGCTG CCGGCCCACT 360

CTTATTTAAG CCGCTTCCCC TGGCGGTCAC AAGACAGAGG CAGGCATG 408

(2) INFORMATION FOR SEQ ID NO:49:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

Ser Pro Thr Ser Pro Glu Pro Pro
1 5

(2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

Asp Thr Thr Ser Pro Glu Ser Pro
1 5

(2) INFORMATION FOR SEQ ID NO:51:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

Asp Thr Thr Ser Pro Glu Pro Pro
1 5

(2) INFORMATION FOR SEQ ID NO:52:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

Asp Lys Thr Ser Pro Glu Pro Ala
1 5

(2) INFORMATION FOR SEQ ID NO:53:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 45 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:

Thr Ser Pro Glu Pro Pro Asp Thr Thr Ser Pro Glu Ser Pro Asp Thr
1 5 10 15

Thr Ser Pro Glu Ser Pro Asp Thr Thr Ser Pro Glu Pro Pro Asp Thr
20 25 30

Thr Ser Pro Glu Pro Pro Asp Lys Thr Ser Pro Glu Pro
35 40 45

(2) INFORMATION FOR SEQ ID NO:54:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

Thr Pro Pro Pro Thr Thr Pro Thr Thr Pro Pro Thr Thr Pro Pro Thr
1 5 10 15

Thr Pro Pro Pro Thr Pro
20

(2) INFORMATION FOR SEQ ID NO:55:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 45 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: not relevant
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

Thr Thr Pro Ser Pro Pro Thr Thr Thr Thr Thr Thr Pro Pro Pro Thr
1 5 10 15

Thr Thr Pro Ser Pro Pro Ile Thr Thr Thr Thr Thr Pro Pro Pro Thr
20 25 30

Thr Thr Pro Ser Pro Pro Ile Ser Thr Thr Thr Thr Pro
35 40 45

(2) INFORMATION FOR SEQ ID NO:56:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 40 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

CGGGGGCCAG GAACTGGAGG CGCCCCCCAG CAGGGCTCCA 40

(2) INFORMATION FOR SEQ ID NO:57:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 40 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

GGGCCCGGAG GTGCAGGAGG CTCCCCGGAG TCTCCCGACA 40

(2) INFORMATION FOR SEQ ID NO:58:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 40 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

GGTGCAGGAG GAGGAGGAGG CTCCCCGGAG CCTCCCGACA 40

(2) INFORMATION FOR SEQ ID NO:59:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 40 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

CTCCCCGGAG CCTCCCGACA CTCCCCGGAG CCTCCCGACA 40

INDICATIONS RELATING TO A DEPOSITED MICROORGANISM

(PCT Rule 13*bis*)

A. The indications made below relate to the microorganism referred to in the description on page 4, line 29.

B. IDENTIFICATION OF DEPOSIT Further deposits are identified on an additional sheet ☐

Name of depositary institution

AMERICAN TYPE CULTURE COLLECTION

Address of depositary institution *(including postal code and country)*

12301 Parklawn Drive
Rockville, Maryland 20852
United States of America

Date of deposit	Accession Number
October 10, 1996	ATCC 97758

C. ADDITIONAL INDICATIONS *(leave blank if not applicable)* This information is continued on an additional sheet ☐

Phage library, PF291

D. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE *(if the indications are not for all designated States)*

E. SEPARATE FURNISHING OF INDICATIONS *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

For receiving Office use only	For International Bureau use only
☒ This sheet was received with the international application	☐ This sheet was received by the International Bureau on:
Authorized officer	Authorized officer

Form PCT/RO/134 (July 1992)

INDICATIONS RELATING TO A DEPOSITED MICROORGANISM

(PCT Rule 13*bis*)

A. The indications made below relate to the microorganism referred to in the description on page 11, line 23.

B. IDENTIFICATION OF DEPOSIT Further deposits are identified on an additional sheet ☐

Name of depositary institution

AMERICAN TYPE CULTURE COLLECTION

Address of depositary institution *(including postal code and country)*

12301 Parklawn Drive
Rockville, Maryland 20852
United States of America

Date of deposit	Accession Number
October 10, 1996	ATCC 97759

C. ADDITIONAL INDICATIONS *(leave blank if not applicable)* This information is continued on an additional sheet ☐

DNA Plasmid, 1321789

D. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE *(if the indications are not for all designated States)*

E. SEPARATE FURNISHING OF INDICATIONS *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

For receiving Office use only	For International Bureau use only
☒ This sheet was received with the international application	☐ This sheet was received by the International Bureau on:
Authorized officer	Authorized officer

Form PCT/RO/134 (July 1992)

What Is Claimed Is:

1. An isolated nucleic acid molecule comprising a polynucleotide having a nucleotide sequence at least 95% identical to a sequence selected from the group consisting of:

(a) a nucleotide sequence encoding the MAdCAM-1 polypeptide having the complete amino acid sequence in FIG. 1 (SEQ ID NO:2), FIG. 2 (SEQ ID NO:4), FIG. 3 (SEQ ID NO:6), FIG. 4 (SEQ ID NO: 8) or FIG. 5 (SEQ ID NO: 10) ;

(b) a nucleotide sequence encoding the mature MAdCAM-1 polypeptide having the amino acid sequence at positions 18-382 in FIG. 1 (SEQ ID NO:2), positions 18-366 in FIG. 2 (SEQ ID NO:4), positions 18-263 in FIG. 3 (SEQ ID NO:6), positions 18-310 in FIG. 4 (SEQ ID NO:8), or positions 18-289 in FIG. 5 (SEQ ID NO:10);

(c) a nucleotide sequence encoding the extracellular domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e));

(d) a nucleotide sequence encoding the intracellular domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e));

(e) a nucleotide sequence encoding the transmembrane domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e));

(f) a nucleotide sequence comprising the MAdCAM-1 promoter, wherein the nucleotide sequence is given in SEQ ID NO:33;

(g) a nucleotide sequence encoding exon 1, 2, 3, 4 or 5 of MAdCAM-1, having the sequence given in SEQ ID NOS:34, 35, 36, 37 and 38, respectively; and

(h) a nucleotide sequence complementary to any of the nucleotide sequences in (a), (b), (c), (d), (e), (f) or (g).

2. The nucleic acid molecule of claim 1 wherein said polynucleotide has the complete nucleotide sequence in FIG. 1 (SEQ ID NO:1), FIG. 2 (SEQ ID NO: 3), FIG. 3 (SEQ ID NO:5), FIG. 4 (SEQ ID NO:7), FIG. 5 (SEQ ID NO:9), SEQ ID NO:33, SEQ ID NO:34, SEQ ID NO:35, SEQ ID NO:36, SEQ ID NO: 37, or SEQ ID NO:38.

3. The nucleic acid molecule of claim 1 wherein said polynucleotide has the nucleotide sequence in FIG. 1 (SEQ ID NO:1) encoding the MAdCAM-1(a) polypeptide having the complete amino acid sequence in FIG. 1 (SEQ ID NO:2), the nucleotide sequence in FIG. 2 (SEQ ID NO:3) encoding the MAdCAM-1(b) polypeptide having the complete amino acid sequence in FIG. 2 (SEQ ID NO:4), the nucleotide sequence in FIG. 3 (SEQ ID NO:5) encoding the MAdCAM-1(c) polypeptide having the complete amino acid sequence in FIG. 3 (SEQ ID NO:6), the nucleotide sequence in FIG. 4 (SEQ ID NO:7) encoding the MAdCAM-1(d) polypeptide having the complete amino acid sequence in FIG. 4 (SEQ ID NO:8), or the nucleotide sequence in FIG. 5 (SEQ ID NO:9) encoding the MAdCAM-1(e) polypeptide having the complete amino acid sequence in FIG. 5 (SEQ ID NO:10).

4. The nucleic acid molecule of claim 1 wherein said polynucleotide has the nucleotide sequence in FIG. 1 (SEQ ID NO:1) encoding the mature MAdCAM-1(a) polypeptide having the amino acid sequence in FIG. 1 (SEQ ID NO:2), the nucleotide sequence in FIG. 2 (SEQ ID NO:3) encoding the mature MAdCAM-1(b) polypeptide having the amino acid sequence in FIG. 2 (SEQ ID NO:4), the nucleotide sequence in FIG. 3 (SEQ ID NO:5) encoding the mature MAdCAM-1(c) polypeptide having the amino acid sequence in FIG. 3 (SEQ ID NO:6), the nucleotide sequence in FIG. 4 (SEQ ID NO:7) encoding the mature MAdCAM-1(d) polypeptide having the amino acid sequence in FIG. 4 (SEQ ID NO:8), or the nucleotide sequence in FIG. 5 (SEQ ID NO:9) encoding the mature

MAdCAM-1(e) polypeptide having the amino acid sequence in FIG. 5 (SEQ ID NO:10).

5. An isolated nucleic acid molecule comprising a polynucleotide which hybridizes under stringent hybridization conditions to a polynucleotide having a nucleotide sequence identical to a nucleotide sequence in (a), (b), (c), (d), (e), (f), or (g) of claim 1 wherein said polynucleotide which hybridizes does not hybridize under stringent hybridization conditions to a polynucleotide having a nucleotide sequence consisting of only A residues or of only T residues.

6. An isolated nucleic acid molecule comprising a polynucleotide which encodes the amino acid sequence of an epitope-bearing portion of any of the MAdCAM-1(a-e) polypeptides having an amino acid sequence in (a), (b), (c), (d), (e) or (g) of claim 1.

7. The isolated nucleic acid molecule of claim 6, which encodes an epitope-bearing portion of any of the MAdCAM-1(a-e) polypeptides selected from the group consisting of: a polypeptide comprising amino acid residues from about 52 to about 80 in FIG. 1 (SEQ ID NO:2); a polypeptide comprising amino acid residues from about 164 to about 196 in FIG. 1 (SEQ ID NO:2); and a polypeptide comprising amino acid residues from about 228 to about 321 in FIG. 1 (SEQ ID NO:2).

8. An isolated nucleic acid molecule comprising a polynucleotide encoding the MAdCAM-1(a) polypeptide having the complete amino acid sequence encoded by the cDNA clone contained in ATCC Deposit No. 97759.

9. An isolated MAdCAM-1(a) polypeptide having an amino acid sequence at least 95% identical to the amino acid sequence of the MAdCAM-1(a) polypeptide having the complete amino acid sequence encoded by the cDNA clone contained in ATCC Deposit No. 97759.

10. An isolated nucleic acid molecule comprising a polynucleotide comprising the MAdCAM-1 promoter having the nucleotide sequence of the genomic clone contained in ATCC Deposit No. 97758.

11. A method for making a recombinant vector comprising inserting an isolated nucleic acid molecule of claim 1 into a vector.

12. A recombinant vector produced by the method of claim 11.

13. A method of making a recombinant host cell comprising introducing the recombinant vector of claim 12 into a host cell.

14. A recombinant host cell produced by the method of claim 13.

15. A recombinant method for producing any of the MAdCAM-1(a-g) polypeptides, comprising culturing the recombinant host cell of claim 14 under conditions such that said polypeptide is expressed and recovering said polypeptide.

16. An isolated MAdCAM-1 polypeptide having an amino acid sequence at least 95% identical to a sequence selected from the group consisting of:

(a) the amino acid sequence of the MAdCAM-1 polypeptide having the complete amino acid sequence in FIG. 1 (SEQ ID NO:2), FIG. 2 (SEQ

ID NO:4), FIG. 3 (SEQ ID NO:6), FIG. 4 (SEQ ID NO: 8) or FIG. 5 (SEQ ID NO: 10);

(b) the amino acid sequence of the mature MAdCAM-1 polypeptide having the amino acid sequence at positions 18-382 in FIG. 1 (SEQ ID NO:2), positions 18-366 in FIG. 2 (SEQ ID NO:4), positions 18-263 in FIG. 3 (SEQ ID NO:6), positions 18-310 in FIG. 4 (SEQ ID NO:8), or positions 18-289 in FIG. 5 (SEQ ID NO:10);

(c) the amino acid sequence of the extracellular domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e));

(d) the amino acid sequence of the intracellular domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e));

(e) the amino acid sequence of the transmembrane domain of any of the MAdCAM-1 polypeptides (MAdCAM-1(a-e));

(f) the amino acid sequence encoded by exon 1, 2, 3, 4 or 5 of MAdCAM-1, wherein said amino acid sequence is encoded by the nucleotide sequence given in SEQ ID NOS:34, 35, 36, 37, 38, or 39, respectively; and

(g) the amino acid sequence of an epitope-bearing portion of any one of the polypeptides of (a), (b), (c), (d), (e) or (f).

17. An isolated polypeptide comprising an epitope-bearing portion of any of the MAdCAM-1 proteins (MAdCAM-1(a-e)), wherein said portion is selected from the group consisting of: a polypeptide comprising amino acid residues from about 52 to about 80 in FIG. 1 (SEQ ID NO:2); a polypeptide comprising amino acid residues from about 164 to about 196 in FIG. 1 (SEQ ID NO:2); and a polypeptide comprising amino acid residues from about 228 to about 321 in FIG. 1 (SEQ ID NO:2).

18. An isolated antibody that binds specifically to a MAdCAM-1 polypeptide of claim 16.

19. A method for treating an individual in need of a reduction in MAdCAM-1(a-e) activity, comprising administering to said individual a therapeutically effective amount of a composition comprising an antagonist of MAdCAM-1(a-e) activity.

20. A method useful during the diagnosis of cancer or of a pathological inflammatory condition, comprising:

(a) assaying the expression level of any of MAdCAM-1(a-e) in mammalian cells or body fluid; and

(b) comparing said expression level of any of MAdCAM-1(a-e) with a standard expression level of any of MAdCAM-1(a-e), whereby an increase in said expression level of any of MAdCAM-1(a-e) over said standard is indicative of cancer or of a pathological inflammatory condition.

21. A recombinant vector comprising a recombinant nucleic acid molecule comprising the 5' flanking region (SEQ ID NO:33), including the promoter, of MAdCAM-1, and a reporter gene, wherein the 5' flanking region is operably linked to the reporter gene.

22. A recombinant host cell comprising the vector of claim 21.

23. A method for the identification of substances capable of altering the expression from the MAdCAM-1 promoter, comprising:

(a) measuring the level of expression of a reporter gene in a test cell, wherein said test cell is transformed with a recombinant DNA molecule comprising a reporter gene operably linked to a DNA molecule comprising the MAdCAM-1 promoter, and wherein a candidate MAdCAM-1 *trans*-acting agent is administered to said test cell;

(b) measuring the level of expression of said reporter gene in a control cell, wherein said control cell is transformed with the recombinant DNA molecule of step (a); and

(c) comparing the level of expression of said reporter gene in said test cell to the level of said reporter gene in said control cell.

```
        10                  30                  50
atggatttcggactggccctcctgctggcggggcttctggggctcctcctcGGCCAGTCC
M  D  F  G  L  A  L  L  L  A  G  L  L  G  L  L  L  G  Q  S
        70                  90                  110
CTCCAGGTGAAGCCCCTGCAGGTGGAGCCCCCGGAGCCGGTGGTGGCCGTGGCCTTGGGC
L  Q  V  K  P  L  Q  V  E  P  P  E  P  V  V  A  V  A  L  G
       130                 150                 170
GCCTCGCGCCAGCTCACCTGCCGCCTGGCCTGCGCGGACCGCGGGGCCTCGGTGCAGTGG
A  S  R  Q  L  T  C  R  L  A  C  A  D  R  G  A  S  V  Q  W
       190                 210                 230
CGGGGCCTGGACACCAGCCTGGGCGCGGTGCAGTCGGACACGGGCCGCAGCGTCCTCACC
R  G  L  D  T  S  L  G  A  V  Q  S  D  T  G  R  S  V  L  T
       250                 270                 290
GTGCGCAACGCCTCGCTGTCGGCGGCCGGGACCCGCGTGTGCGTGGGCTCCTGCGGGGGC
V  R  N  A  S  L  S  A  A  G  T  R  V  C  V  G  S  C  G  G
       310                 330                 350
CGCACCTTCCAGCACACCGTGCAGCTCCTTGTGTACGCCTTCCCGGACCAGCTGACCGTC
R  T  F  Q  H  T  V  Q  L  L  V  Y  A  F  P  D  Q  L  T  V
       370                 390                 410
TCCCCAGCAGCCCTGGTGCCTGGTGACCCGGAGGTGGCCTGTACGGCCCACAAAGTCACG
S  P  A  A  L  V  P  G  D  P  E  V  A  C  T  A  H  K  V  T
       430                 450                 470
CCCGTGGACCCCAACGCGCTCTCCTTCTCCCTGCTCGTCGGGGGCCAGGAACTGGAGGGG
P  V  D  P  N  A  L  S  F  S  L  L  V  G  G  Q  E  L  E  G
       490                 510                 530
GCGCAAGCCCTGGGCCCGGAGGTGCAGGAGGAGGAGGAGGAGCCCCAGGGGGACGAGGAC
A  Q  A  L  G  P  E  V  Q  E  E  E  E  E  P  Q  G  D  E  D
       550                 570                 590
GTGCTGTTCAGGGTGACAGAGCGCTGGCGGCTGCCGCCCCTGGGGACCCCTGTCCCGCCC
V  L  F  R  V  T  E  R  W  R  L  P  P  L  G  T  P  V  P  P
       610                 630                 650
GCCCTCTACTGCCAGGCCACGATGAGGCTGCCTGGCTTGGAGCTCAGCCACCGCCAGGCC
A  L  Y  C  Q  A  T  M  R  L  P  G  L  E  L  S  H  R  Q  A
       670                 690                 710
ATCCCCGTCCTGCACAGCCCGACCTCCCCGGAGCCTCCCGACACCACCTCCCCGGAGTCT
I  P  V  L  H  S  P  T  S  P  E  P  P  D  T  T  S  P  E  S
       730                 750                 770
CCCGACACCACCTCCCCGGAGTCTCCCGACACCACCTCCCCGGAGCCTCCCGACACCACC
P  D  T  T  S  P  E  S  P  D  T  T  S  P  E  P  P  D  T  T
       790                 810                 830
TCCCCGGAGCCTCCCGACAAGACCTCCCCGGAGCCCGCCCCCCAGCAGGGCTCCACACAC
S  P  E  P  P  D  K  T  S  P  E  P  A  P  Q  Q  G  S  T  H
```

FIG.1A

```
          850                 870                 890
ACCCCCAGGAGCCCAGGCTCCACCAGGACTCGCCGCCCTGAGATCTCCCAGGCTGGGCCC
T  P  R  S  P  G  S  T  R  T  R  R  P  E  I  S  Q  A  G  P
          910                 930                 950
ACGCAGGGAGAAGTGATCCCAACAGGCTCGTCCAAACCTGCGGGTGACCAGCTGCCCGCG
T  Q  G  E  V  I  P  T  G  S  S  K  P  A  G  D  Q  L  P  A
          970                 990                1010
GCTCTGTGGACCAGCAGTGCGGTGCTGGGACTGCTGCTCCTGGCCTTGCCCACGTATCAC
A  L  W  T  S  S  A  V  L  G  L  L  L  L  A  L  P  T  Y  H
         1030                1050                1070
CTCTGGAAACGCTGCCGGCACCTGGCTGAGGACGACACCCACCCACCAGCTTCTCTGAGG
L  W  K  R  C  R  H  L  A  E  D  D  T  H  P  P  A  S  L  R
         1090                1110                1130
CTTCTGCCCCAGGTGTCGGCCTGGGCTGGGTTAAGGGGGACCGGCCAGGTCGGGATCAGC
L  L  P  Q  V  S  A  W  A  G  L  R  G  T  G  Q  V  G  I  S
         1150                1170                1190
CCCTCCTGAGTGGCCAGCCTTTCCCCCTGTGAAAGCAAAATAGCTTGGACCCCTTCAAGT
P  S  *
         1210                1230                1250
TGAGAACTGGTCAGGGCAAACCTGCCTCCCATTCTACTCAAAGTCATCCCTCTGTTCACA
         1270                1290                1310
GAGATGGATGCATGTTCTGATTGCCTCTTTGGAGAAGCTCATCAGAAACTCAAAAGAAGG
         1330                1350                1370
CCACTGTTTGTCTCACCTACCCATGACCTGAAGCCCCTCCCTGAGTGGTCCCCACCTTTC
         1390                1410                1430
TGGACGGAACCACGTACTTTTTACATACATTGATTCATGTCTCACGTCTCCCTAAAAATG
         1450                1470                1490
CGTAAGACCAAGCTGTGCCCTGACCACCCTGGGCCCCTGTCGTCAGGACCTCCTGAGGCT
         1510                1530
TTGGCAAATAAACCTCCTAAAATGATAAAAAAAAAA
```

FIG.1B

```
        10                  30                  50
atggatttcggactggccctcctgctggcggggcttctggggctcctcctcGGCCAGTCC
M D F G L A L L L A G L L G L L L G Q S
        70                  90                  110
CTCCAGGTGAAGCCCCTGCAGGTGGAGCCCCCGGAGCCGGTGGTGGCCGTGGCCTTGGGC
L Q V K P L Q V E P P E P V V A V A L G
        130                 150                 170
GCCTCGCGCCAGCTCACCTGCCGCCTGGCCTGCGCGGACCGCGGGGCCTCGGTGCAGTGG
A S R Q L T C R L A C A D R G A S V Q W
        190                 210                 230
CGGGGCCTGGACACCAGCCTGGGCGCGGTGCAGTCGGACACGGGCCGCAGCGTCCTCACC
R G L D T S L G A V Q S D T G R S V L T
        250                 270                 290
GTGCGCAACGCCTCGCTGTCGGCGGCCGGGACCCGCGTGTGCGTGGGCTCCTGCGGGGGC
V R N A S L S A A G T R V C V G S C G G
        310                 330                 350
CGCACCTTCCAGCACACCGTGCAGCTCCTTGTGTACGCCTTCCCGGACCAGCTGACCGTC
R T F Q H T V Q L L V Y A F P D Q L T V
        370                 390                 410
TCCCCAGCAGCCCTGGTGCCTGGTGACCCGGAGGTGGCCTGTACGGCCCACAAAGTCACG
S P A A L V P G D P E V A C T A H K V T
        430                 450                 470
CCCGTGGACCCCAACGCGCTCTCCTTCTCCCTGCTCGTCGGGGGCCAGGAACTGGAGGGG
P V D P N A L S F S L L V G G Q E L E G
        490                 510                 530
GCGCAAGCCCTGGGCCCGGAGGTGCAGGAGGAGGAGGAGGAGCCCCAGGGGGACGAGGAC
A Q A L G P E V Q E E E E E P Q G D E D
        550                 570                 590
GTGCTGTTCAGGGTGACAGAGCGCTGGCGGCTGCCGCCCCTGGGGACCCCTGTCCCGCCC
V L F R V T E R W R L P P L G T P V P P
        610                 630                 650
GCCCTCTACTGCCAGGCCACGATGAGGCTGCCTGGCTTGGAGCTCAGCCACCGCCAGGCC
A L Y C Q A T M R L P G L E L S H R Q A
        670                 690                 710
ATCCCCGTCCTGCACAGCCCGACCTCCCCGGAGTCTCCCGACACCACCTCCCCGGAGCCT
I P V L H S P T S P E S P D T T S P E P
        730                 750                 770
CCCGACACCACCTCCCCGGAGCCTCCCGACAAGACCTCCCCGGAGCCCGCCCCCCAGCAG
P D T T S P E P P D K T S P E P A P Q Q
        790                 810                 830
GGCTCCACACACACCCCCAGGAGCCCAGGCTCCACCAGGACTCGCCGCCCTGAGATCTCC
G S T H T P R S P G S T R T R R P E I S
```

FIG.2A

```
      850                 870                 890
CAGGCTGGGCCCACGCAGGGAGAAGTGATCCCAACAGGCTCGTCCAAACCTGCGGGTGAC
Q  A  G  P  T  Q  G  E  V  I  P  T  G  S  S  K  P  A  G  D
      910                 930                 950
CAGCTGCCCGCGGCTCTGTGGACCAGCAGTGCGGTGCTGGGACTGCTGCTCCTGGCCTTG
Q  L  P  A  A  L  W  T  S  S  A  V  L  G  L  L  L  L  A  L
      970                 990                 1010
CCCACGTATCACCTCTGGAAACGCTGCCGGCACCTGGCTGAGGACGACACCCACCCACCA
P  T  Y  H  L  W  K  R  C  R  H  L  A  E  D  D  T  H  P  P
      1030                1050                1070
GCTTCTCTGAGGCTTCTGCCCCAGGTGTCGGCCTGGGCTGGGTTAAGGGGGACCGGCCAG
A  S  L  R  L  L  P  Q  V  S  A  W  A  G  L  R  G  T  G  Q
      1090                1110                1130
GTCGGGATCAGCCCCTCCTGAGTGGCCAGCCTTTCCCCCTGTGAAAGCAAAATAGCTTGG
V  G  I  S  P  S  *
      1150                1170                1190
ACCCCTTCAAGTTGAGAACTGGTCAGGGCAAACCTGCCTCCCATTCTACTCAAAGTCATC
      1210                1230                1250
CCTCTGTTCACAGAGATGGATGCATGTTCTGATTGCCTCTTTGGAGAAGCTCATCAGAAA
      1270                1290                1310
CTCAAAAGAAGGCCACTGTTTGTCTCACCTACCCATGACCTGAAGCCCCTCCCTGAGTGG
      1330                1350                1370
TCCCCACCTTTCTGGACGGAACCACGTACTTTTTACATACATTGATTCATGTCTCACGTC
      1390                1410                1430
TCCCTAAAAATGCGTAAGACCAAGCTGTGCCCTGACCACCCTGGGCCCCTGTCGTCAGGA
      1450                1470
CCTCCTGAGGCTTTGGCAAATAAACCTCCTAAAATGATAAAAAAAAAA
```

FIG.2B

```
        10                  30                  50
ATGGATTTCGGACTGGCCCTCCTGCTGGCGGGGCTTCTGGGGCTCCTCCTCGGCCAGTCC
M  D  F  G  L  A  L  L  L  A  G  L  L  G  L  L  L  G  Q  S
        70                  90                  110
CTCCAGGTGAAGCCCCTGCAGGTGGAGCCCCCGGAGCCGGTGGTGGCCGTGGCCTTGGGC
L  Q  V  K  P  L  Q  V  E  P  P  E  P  V  V  A  V  A  L  G
        130                 150                 170
GCCTCGCGCCAGCTCACCTGCCGCCTGGCCTGCGCGGACCGCGGGGCCTCGGTGCAGTGG
A  S  R  Q  L  T  C  R  L  A  C  A  D  R  G  A  S  V  Q  W
        190                 210                 230
CGGGGCCTGGACACCAGCCTGGGCGCGGTGCAGTCGGACACGGGCCGCAGCGTCCTCACC
R  G  L  D  T  S  L  G  A  V  Q  S  D  T  G  R  S  V  L  T
        250                 270                 290
GTGCGCAACGCCTCGCTGTCGGCGGCCGGGACCCGCGTGTGCGTGGGCTCCTGCGGGGGC
V  R  N  A  S  L  S  A  A  G  T  R  V  C  V  G  S  C  G  G
        310                 330                 350
CGCACCTTCCAGCACACCGTGCAGCTCCTTGTGTACGCCTTCCCGGACCAGCTGACCGTC
R  T  F  Q  H  T  V  Q  L  L  V  Y  A  F  P  D  Q  L  T  V
        370                 390                 410
TCCCCAGCAGCCCTGGTGCCTGGTGACCCGGAGGTGGCCTGTACGGCCCACAAAGTCACG
S  P  A  A  L  V  P  G  D  P  E  V  A  C  T  A  H  K  V  T
        430                 450                 470
CCCGTGGACCCCAACGCGCTCTCCTTCTCCCTGCTCGTCGGGGGCCAGCAGGGCTCCACA
P  V  D  P  N  A  L  S  F  S  L  L  V  G  G  Q  Q  G  S  T
        490                 510                 530
CACACCCCCAGGAGCCCAGGCTCCACCAGGACTCGCCGCCCTGAGATCTCCCAGGCTGGG
H  T  P  R  S  P  G  S  T  R  T  R  R  P  E  I  S  Q  A  G
        550                 570                 590
CCCACGCAGGGAGAAGTGATCCCAACAGGCTCGTCCAAACCTGCGGGTGACCAGCTGCCC
P  T  Q  G  E  V  I  P  T  G  S  S  K  P  A  G  D  Q  L  P
        610                 630                 650
GCGGCTCTGTGGACCAGCAGTGCGGTGCTGGGACTGCTGCTCCTGGCCTTGCCCACGTAT
A  A  L  W  T  S  S  A  V  L  G  L  L  L  L  A  L  P  T  Y
        670                 690                 710
CACCTCTGGAAACGCTGCCGGCACCTGGCTGAGGACGACACCCACCCACCAGCTTCTCTG
H  L  W  K  R  C  R  H  L  A  E  D  D  T  H  P  P  A  S  L
        730                 750                 770
AGGCTTCTGCCCCAGGTGTCGGCCTGGGCTGGGTTAAGGGGGACCGGCCAGGTCGGGATC
R  L  L  P  Q  V  S  A  W  A  G  L  R  G  T  G  Q  V  G  I
        790                 810                 830
AGCCCCTCCTGAGTGGCCAGCCTTTCCCCCTGTGAAAGCAAAATAGCTTGGACCCCTTCA
S  P  S  *
```

FIG.3A

```
       850                 870                 890
AGTTGAGAACTGGTCAGGGCAAACCTGCCTCCCATTCTACTCAAAGTCATCCCTCTGTTC
       910                 930                 950
ACAGAGATGGATGCATGTTCTGATTGCCTCTTTGGAGAAGCTCATCAGAAACTCAAAAGA
       970                 990                1010
AGGCCACTGTTTGTCTCACCTACCCATGACCTGAAGCCCCTCCCTGAGTGGTCCCCACCT
      1030                1050                1070
TTCTGGACGGAACCACGTACTTTTTACATACATTGATTCATGTCTCACGTCTCCCTAAAA
      1090                1110                1130
ATGCGTAAGACCAAGCTGTGCCCTGACCACCCTGGGCCCCTGTCGTCAGGACCTCCTGAG
      1150                1170
GCTTTGGCAAATAAACCTCCTAAAATGATAAAAAAAAAA
```

FIG.3B

```
        10                  30                  50
ATGGATTTCGGACTGGCCCTCCTGCTGGCGGGGCTTCTGGGGCTCCTCCTCGGCCAGTCC
M  D  F  G  L  A  L  L  L  A  G  L  L  G  L  L  L  G  Q  S
        70                  90                  110
CTCCAGGTGAAGCCCCTGCAGGTGGAGCCCCCGGAGCCGGTGGTGGCCGTGGCCTTGGGC
L  Q  V  K  P  L  Q  V  E  P  P  E  P  V  V  A  V  A  L  G
        130                 150                 170
GCCTCGCGCCAGCTCACCTGCCGCCTGGCCTGCGCGGACCGCGGGGCCTCGGTGCAGTGG
A  S  R  Q  L  T  C  R  L  A  C  A  D  R  G  A  S  V  Q  W
        190                 210                 230
CGGGGCCTGGACACCAGCCTGGGCGCGGTGCAGTCGGACACGGGCCGCAGCGTCCTCACC
R  G  L  D  T  S  L  G  A  V  Q  S  D  T  G  R  S  V  L  T
        250                 270                 290
GTGCGCAACGCCTCGCTGTCGGCGGCCGGGACCCGCGTGTGCGTGGGCTCCTGCGGGGGC
V  R  N  A  S  L  S  A  A  G  T  R  V  C  V  G  S  C  G  G
        310                 330                 350
CGCACCTTCCAGCACACCGTGCAGCTCCTTGTGTACGCCTTCCCGGACCAGCTGACCGTC
R  T  F  Q  H  T  V  Q  L  L  V  Y  A  F  P  D  Q  L  T  V
        370                 390                 410
TCCCCAGCAGCCCTGGTGCCTGGTGACCCGGAGGTGGCCTGTACGGCCCACAAAGTCACG
S  P  A  A  L  V  P  G  D  P  E  V  A  C  T  A  H  K  V  T
        430                 450                 470
CCCGTGGACCCCAACGCGCTCTCCTTCTCCCTGCTCGTCGGGGGCCAGGAACTGGAGGGG
P  V  D  P  N  A  L  S  F  S  L  L  V  G  G  Q  E  L  E  G
        490                 510                 530
GCGCAAGCCCTGGGCCCGGAGTCTCCCGACACCACCTCCCCGGAGTCTCCCGACACCACC
A  Q  A  L  G  P  E  S  P  D  T  T  S  P  E  S  P  D  T  T
        550                 570                 590
TCCCCGGAGCCTCCCGACACCACCTCCCCGGAGCCTCCCGACAAGACCTCCCCGGAGCCC
S  P  E  P  P  D  T  T  S  P  E  P  P  D  K  T  S  P  E  P
        610                 630                 650
GCCCCCCAGCAGGGCTCCACACACACCCCCAGGAGCCCAGGCTCCACCAGGACTCGCCGC
A  P  Q  Q  G  S  T  H  T  P  R  S  P  G  S  T  R  T  R  R
        670                 690                 710
CCTGAGATCTCCCAGGCTGGGCCCACGCAGGGAGAAGTGATCCCAACAGGCTCGTCCAAA
P  E  I  S  Q  A  G  P  T  Q  G  E  V  I  P  T  G  S  S  K
        730                 750                 770
CCTGCGGGTGACCAGCTGCCCGCGGCTCTGTGGACCAGCAGTGCGGTGCTGGGACTGCTG
P  A  G  D  Q  L  P  A  A  L  W  T  S  S  A  V  L  G  L  L
        790                 810                 830
CTCCTGGCCTTGCCCACGTATCACCTCTGGAAACGCTGCCGGCACCTGGCTGAGGACGAC
L  L  A  L  P  T  Y  H  L  W  K  R  C  R  H  L  A  E  D  D
```

FIG.4A

```
       850                 870                 890
ACCCACCCACCAGCTTCTCTGAGGCTTCTGCCCCAGGTGTCGGCCTGGGCTGGGTTAAGG
T  H  P  P  A  S  L  R  L  L  P  Q  V  S  A  W  A  G  L  R
       910                 930                 950
GGGACCGGCCAGGTCGGGATCAGCCCCTCCTGAGTGGCCAGCCTTTCCCCCTGTGAAAGC
G  T  G  Q  V  G  I  S  P  S  *
       970                 990                1010
AAAATAGCTTGGACCCCTTCAAGTTGAGAACTGGTCAGGGCAAACCTGCCTCCCATTCTA
      1030                1050                1070
CTCAAAGTCATCCCTCTGTTCACAGAGATGGATGCATGTTCTGATTGCCTCTTTGGAGAA
      1090                1110                1130
GCTCATCAGAAACTCAAAAGAAGGCCACTGTTTGTCTCACCTACCCATGACCTGAAGCCC
      1150                1170                1190
CTCCCTGAGTGGTCCCCACCTTTCTGGACGGAACCACGTACTTTTTACATACATTGATTC
      1210                1230                1250
ATGTCTCACGTCTCCCTAAAAATGCGTAAGACCAAGCTGTGCCCTGACCACCCTGGGCCC
      1270                1290                1310
CTGTCGTCAGGACCTCCTGAGGCTTTGGCAAATAAACCTCCTAAAATGATAAAAAAAAAA
```

FIG.4B

```
        10                  30                  50
ATGGATTTCGGACTGGCCCTCCTGCTGGCGGGGCTTCTGGGGCTCCTCCTCGGCCAGTCC
M  D  F  G  L  A  L  L  L  A  G  L  L  G  L  L  L  G  Q  S
        70                  90                  110
CTCCAGGTGAAGCCCCTGCAGGTGGAGCCCCCGGAGCCGGTGGTGGCCGTGGCCTTGGGC
L  Q  V  K  P  L  Q  V  E  P  P  E  P  V  V  A  V  A  L  G
        130                 150                 170
GCCTCGCGCCAGCTCACCTGCCGCCTGGCCTGCGCGGACCGCGGGGCCTCGGTGCAGTGG
A  S  R  Q  L  T  C  R  L  A  C  A  D  R  G  A  S  V  Q  W
        190                 210                 230
CGGGGCCTGGACACCAGCCTGGGCGCGGTGCAGTCGGACACGGGCCGCAGCGTCCTCACC
R  G  L  D  T  S  L  G  A  V  Q  S  D  T  G  R  S  V  L  T
        250                 270                 290
GTGCGCAACGCCTCGCTGTCGGCGGCCGGGACCCGCGTGTGCGTGGGCTCCTGCGGGGGC
V  R  N  A  S  L  S  A  A  G  T  R  V  C  V  G  S  C  G  G
        310                 330                 350
CGCACCTTCCAGCACACCGTGCAGCTCCTTGTGTACGCCTTCCCGGACCAGCTGACCGTC
R  T  F  Q  H  T  V  Q  L  L  V  Y  A  F  P  D  Q  L  T  V
        370                 390                 410
TCCCCAGCAGCCCTGGTGCCTGGTGACCCGGAGGTGGCCTGTACGGCCCACAAAGTCACG
S  P  A  A  L  V  P  G  D  P  E  V  A  C  T  A  H  K  V  T
        430                 450                 470
CCCGTGGACCCCAACGCGCTCTCCTTCTCCCTGCTCGTCGGGGGCCAGGAACTGGAGGGG
P  V  D  P  N  A  L  S  F  S  L  L  V  G  G  Q  E  L  E  G
        490                 510                 530
GCGCAAGCCCTGGGCCCGGAGGTGCAGGAGTCTCCCGACACCACCTCCCCGGAGTCTCCC
A  Q  A  L  G  P  E  V  Q  E  S  P  D  T  T  S  P  E  S  P
        550                 570                 590
GACACCACCTCCCCGGAGCCTCCCGACACCACCTCCCCGGAGCCTCCCGACAAGACCTCC
D  T  T  S  P  E  P  P  D  T  T  S  P  E  P  P  D  K  T  S
        610                 630                 650
CCGGAGCCCGCCCCCCAGCAGGGCTCCACACACACCCCCAGGAGCCCAGGCTCCACCAGG
P  E  P  A  P  Q  Q  G  S  T  H  T  P  R  S  P  G  S  T  R
        670                 690                 710
ACTCGCCGCCCTGAGATCTCCCAGGCTGGGCCCACGCAGGGAGAAGTGATCCCAACAGGC
T  R  R  P  E  I  S  Q  A  G  P  T  Q  G  E  V  I  P  T  G
        730                 750                 770
TCGTCCAAACCTGCGGGTGACCAGCTGCCCGCGGCTCTGTGGACCAGCAGTGCGGTGCTG
S  S  K  P  A  G  D  Q  L  P  A  A  L  W  T  S  S  A  V  L
        790                 810                 830
GGACTGCTGCTCCTGGCCTTGCCCACGTATCACCTCTGGAAACGCTGCCGGCACCTGGCT
G  L  L  L  L  A  L  P  T  Y  H  L  W  K  R  C  R  H  L  A
```

FIG.5A

```
       850                 870                 890
GAGGACGACACCCACCCACCAGCTTCTCTGAGGCTTCTGCCCCAGGTGTCGGCCTGGGCT
E  D  D  T  H  P  P  A  S  L  R  L  L  P  Q  V  S  A  W  A
       910                 930                 950
GGGTTAAGGGGGACCGGCCAGGTCGGGATCAGCCCCTCCTGAGTGGCCAGCCTTTCCCCC
G  L  R  G  T  G  Q  V  G  I  S  P  S  *
       970                 990                1010
TGTGAAAGCAAAATAGCTTGGACCCCTTCAAGTTGAGAACTGGTCAGGGCAAACCTGCCT
      1030                1050                1070
CCCATTCTACTCAAAGTCATCCCTCTGTTCACAGAGATGGATGCATGTTCTGATTGCCTC
      1090                1110                1130
TTTGGAGAAGCTCATCAGAAACTCAAAAGAAGGCCACTGTTTGTCTCACCTACCCATGAC
      1150                1170                1190
CTGAAGCCCCTCCCTGAGTGGTCCCCACCTTTCTGGACGGAACCACGTACTTTTTACATA
      1210                1230                1250
CATTGATTCATGTCTCACGTCTCCCTAAAAATGCGTAAGACCAAGCTGTGCCCTGACCAC
      1270                1290                1310
CCTGGGCCCCTGTCGTCAGGACCTCCTGAGGCTTTGGCAAATAAACCTCCTAAAATGATA

AAAAAAAAA
```

FIG.5B

MAdCAM-1 5' FLANKING REGION

1 CTGCAGCTCC GGAACGGGGG GGGGCTGCTC TCCACCGCCC CTGTGCGGCC

51 GCCCGGGAAA GTGCAGGCGG GCCGGGCGCG GTGGCTCACG CCTGTGATCT

101 CAGCACTTTG GGAGGCCGAG GTGGGCGGAT CACCTGAGGT CGGGAGTTCG

151 AGGCCAGCCT GCCCAACATG GAGAAACCCT GTCTCTACTA AAGATACAAA

201 ATTAGCCAGG CGTGGTGACG CATGCCTGTA ATCCCAGCTA CTGGAGTGGC

251 TGAGGCAGGA GAATCGCTTG AGCCCGGGAG ACAGAGGTTG CGGTGAGCTG

301 AGATCGCACC ATTGCAACTC CAGCCTGGGC AACAAGAGCG AAACTCAGAA

351 AAAAAAGAAA AGAAAGTGCA GGGGACCCGC CGTCGGGGTG GGGGCGGCGC

401 TGCCCAGCCT CTGTCCCACT TCCATGCACT TGACCTCGAC CCTCCGGCCT

451 CCGTCTGCGA TCTTCCCGTG CCTGAATATG AGGCTTGGAA CAGACCCAGA

501 CCTTCCTGCC TGCCCGTCCT GAGTGGCCCC GGGACCCCGC CCCATCTTTG

551 GCCCCCAGCC CCTGCCTTTT TGCCGCCTCC AGGGTCGGGG GTCAGGCCAG

601 GAAAGCCCCT TGGGAAGCCC CCGGGGAGCA GCTGGAGCGG GGTCGCCGGG

651 CGGCGGGAAG GAGTGGGCGC CTCTATTTAA GCGGCTTCCC CGCGGCCTCG

701 GGACAGAGGG GACTGAGC

FIG.6A

EXON 1
.....ATGGATTTCG GACTGGCCCT CCTGCTGGCG GGGCTTCTGG GGCTCCTCCT CGgtgagaag 60
gg..... 62

EXON 2
.....gtcgccgcag GCCAGTCCCT CCAGGTGAAG CCCCTGCAGG TGGAGCCCCC GGAGCCGGTG 60
GTGGCCGTGG CCTTGGGCGC CTCGCGCCAG CTCACCTGCC GCCTGGCCTG CGCGGACCGC 120
GGGGCCTCGG TGCAGTGGCG GGGCCTGGAC ACCAGCCTGG GCGCGGTGCA GTCGGACACG 180
GGCCGCAGCG TCCTCACCGT GCGCAACGCC TCGCTGTCGG CGGCCGGGAC CCGCGTGTGC 240
GTGGGCTCCT GCGGGGGCCG CACCTTCCAG CACACCGTGC AGCTCCTTGT GTACGgtgag 300
gcgtc..... 305

EXON 3
.....tccatcacag CCTTCCCGGA CCAGCTGACC GTCTCCCCAG CAGCCCTGGT GCCTGGTGAC 60
CCGGAGGTGG CCTGTACGGC CCACAAAGTC ACGCCCGTGG ACCCCAACGC GCTCTCCTTC 120
TCCCTGCTCG TCGGGGGCCA GGAACTGGAG GGGGCGCAAG CCCTGGGCCC GGAGGTGCAG 180
GAGGAGGAGG AGGAGCCCCA GGGGGACGAG GACGTGCTGT TCAGGGTGAC AGAGCGCTGG 240
CGGCTGCCGC CCCTGGGGAC CCCTGTCCCG CCCGCCCTCT ACTGCCAGGC CACGATGAGG 300
CTGCCTGGCT TGGAGCTCAG CCACCGCCAG GCCATCCCCG gtgagtccgc..... 350

EXON 4
.....ctgtttccag TCCTGCACAG CCCGACCTCC CCGGAGCCTC CCGACACCAC CTCCCCGGAG 60
CCTCCCAACA CCACCTCCCC GGAGTCTCCC GACACCACCT CCCCGGAGTC TCCCGACACC 120
ACCTCCCAGG AGCCTCCCGA CACCACCTCC CAGGAGCCTC CCGACACCAC CTCCCAGGAG 180
CCTCCCGACA CCACCTCCCC GGAGCCTCCC GACAAGACCT CCCCGGAGCC CGCCCCCCAG 240
CAGGGCTCCA CACACACCCC CAGGAGCCCA GGCTCCACCA GGACTCGCCG CCCTGAGATC 300
TCCCAGGCTG GGCCCACGCA GGGAGAAGTG ATCCCAACAG GCTgtgagtt ctg..... 353

EXON 5
.....ctctccccag CGTCCAAACC TGCGGGTGAC CAGCTGCCCG CGGCTCTGTG GACCAGCAGT 60
GCGGTGCTGG GACTGCTGCT CCTGGCCTTG CCCACCTATC ACCTCTGGAA ACGCTGCCGG 120
CACCTGGCTG AGGACGACAC CCACCCACCA GCTTCTCTGA GGCTTCTGCC CCAGGTGTCG 180
GCCTGGGCTG GGTTAAGGGG GACCGGCCAG GTCGGGATCA GCCCCTCCTG AGTGGCCAGC 240
CTTTCCCCCT GTGAAAGCAA AATAGCTTGG ACCCCTTCAA GTTGAGAACT GGTCAGGGCA 300
AACCTGCCTC CCATTCTACT CAAAGTCATC CCTCTGTTCA CAGAGATGGA TGCATGTTCT 360
GATTGCCTCT TTGGAGAAGC TCATCAGAAA CTCAAAAGAA GGCCACTGTT TGTCTCACCT 420
ACCCATGACC TGAAGCCCCT CCCTGAGTGG TCCCCACCTT TCTGGACGGA ACCACGTACT 480
TTTTACATAC ATTGATTCAT GTCTCACGTC TCCCTAAAAA TGCGTAAGAC CAAGCTGTGC 540
CCTGACCACC CTGGGCCCCT GTCGTCAGGA CCTCCTGAGG CTTTGGCAAA TAAACCTCCT 600
AAAATGAT 608

FIG.6B

```
1   MESILALLLA--LALVPYQLSRGQSFQVNPPESEVAVAMG  MAdCAM-1 mouse
1   MDFGLALLLAGLLLGQSLQVKPLQVEPPEPVVAVALG  MAdCAM-1
1   MDFGLALLLAGLLLGQSLQVKPLQVEPPEPVVAVALG  MAdCAM-1 a
1   MDFGLALLLAGLLLGQSLQVKPLQVEPPEPVVAVALG  MAdCAM-1 b
1   MDFGLALLLAGLLLGQSLQVKPLQVEPPEPVVAVALG  MAdCAM-1 c
1   MDFGLALLLAGLLLGQSLQVKPLQVEPPEPVVAVALG  MAdCAM-1 d
1   MDFGLALLLAGLLLGQSLQVKPLQVEPPEPVVAVALG  MAdCAM-1 e

39  ISLQITCSMSCDEGVARVHWRGLDTSLGSVQTLPGSSILS  MAdCAM-1 mouse
41  ASRQLTCRLACADRGASVQWRGLDTSLGAVQSDTGRSVLT  MAdCAM-1
41  ASRQLTCRLACADRGASVQWRGLDTSLGAVQSDTGRSVLT  MAdCAM-1 a
41  ASRQLTCRLACADRGASVQWRGLDTSLGAVQSDTGRSVLT  MAdCAM-1 b
41  ASRQLTCRLACADRGASVQWRGLDTSLGAVQSDTGRSVLT  MAdCAM-1 c
41  ASRQLTCRLACADRGASVQWRGLDTSLGAVQSDTGRSVLT  MAdCAM-1 d
41  ASRQLTCRLACADRGASVQWRGLDTSLGAVQSDTGRSVLT  MAdCAM-1 e

79  VRGM-LSDIGTPVCVGSCGSRSFQHSVKILVYAFPDQLVV  MAdCAM-1 mouse
81  VRNASLSAAGTRVCVGSCGGRTFQHTVQLLVYAFPDQLTV  MAdCAM-1
81  VRNASLSAAGTRVCVGSCGGRTFQHTVQLLVYAFPDQLTV  MAdCAM-1 a
81  VRNASLSAAGTRVCVGSCGGRTFQHTVQLLVYAFPDQLTV  MAdCAM-1 b
81  VRNASLSAAGTRVCVGSCGGRTFQHTVQLLVYAFPDQLTV  MAdCAM-1 c
81  VRNASLSAAGTRVCVGSCGGRTFQHTVQLLVYAFPDQLTV  MAdCAM-1 d
81  VRNASLSAAGTRVCVGSCGGRTFQHTVQLLVYAFPDQLTV  MAdCAM-1 e
```

FIG.7A

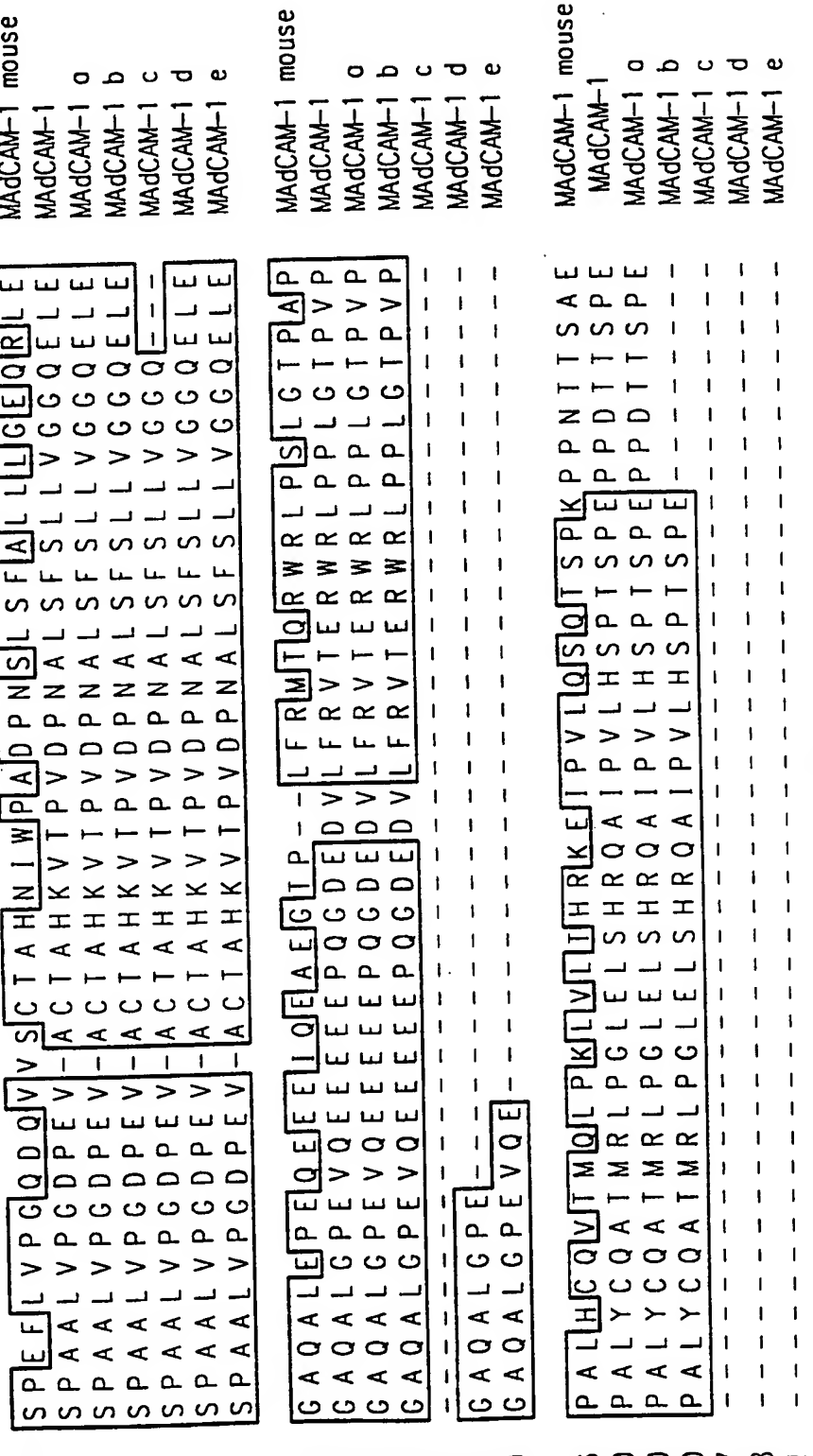
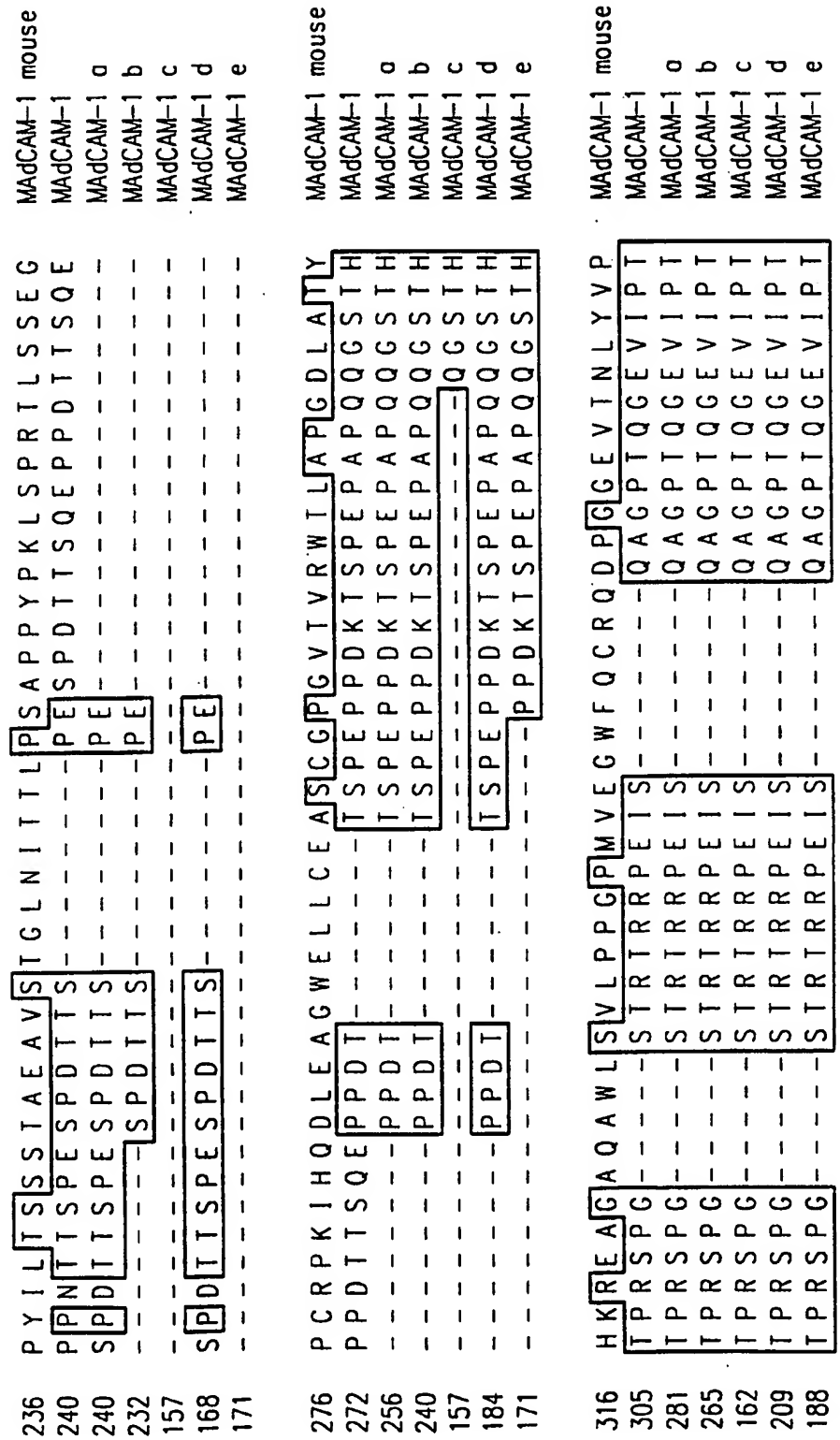

MAdCAM-1 mouse
MAdCAM-1
MAdCAM-1 a
MAdCAM-1 b
MAdCAM-1 c
MAdCAM-1 d
MAdCAM-1 e
118
121
121
121
121
121
121
158
160
160
160
157
160
160
196
200
200
200
157
168
171

FIG.7B


236 P Y I L T S S S T A E A V S T G L N I T T L P S A P P Y P K L S P R T L S S E G MAdCAM-1 mouse
240 P P N T T S P E S P D T T S - - - - - - - - - - - - - - - - P E S P D T T S Q E P P D T T S Q E MAdCAM-1
240 S P D T T S P E S P D T T S - - - - - - - - - - - - - - - - P E - - - - - - - - - - - - - - - MAdCAM-1 a
232 - - - - - - - - - - - - - - - - S P D T T S - - - - - - - - - - - - - - - - P E - - - - - - - - - - - - - - - MAdCAM-1 b
157 - MAdCAM-1 c
168 S P D T T S P E S P D T T S - - - - - - - - - - - - - - - - P E - - - - - - - - - - - - - - - MAdCAM-1 d
171 - MAdCAM-1 e

276 P C R P K I H Q D L E A G W E L L C E A S C G P G V T V R W T L A P G D L A T Y MAdCAM-1 mouse
272 P P D T T S Q E P P D T - - - - - - - - - - T S P E P P D K T S P E P A P Q Q G S T H MAdCAM-1
256 - - - - - - - - - - - - - - - - P P D T - - - - - - - - - - T S P E P P D K T S P E P A P Q Q G S T H MAdCAM-1 a
240 - - - - - - - - - - - - - - - - P P D T - - - - - - - - - - T S P E P P D K T S P E P A P Q Q G S T H MAdCAM-1 b
157 - MAdCAM-1 c
184 - - - - - - - - - - - - - - - - P P D T - - - - - - - - - - T S P E P P D K T S P E P A P Q Q G S T H MAdCAM-1 d
171 - P P D K T S P E P A P Q Q G S T H MAdCAM-1 e

316 H K R E A G A Q A W L S V L P P G P M V E G W F Q C R Q D P G G E V T N L Y V P MAdCAM-1 mouse
305 T P R S P G - - - - - - S T R T R R P E I S - - - - - - - - - - Q A G P T Q G E V I P T MAdCAM-1
281 T P R S P G - - - - - - S T R T R R P E I S - - - - - - - - - - Q A G P T Q G E V I P T MAdCAM-1 a
265 T P R S P G - - - - - - S T R T R R P E I S - - - - - - - - - - Q A G P T Q G E V I P T MAdCAM-1 b
162 T P R S P G - - - - - - S T R T R R P E I S - - - - - - - - - - Q A G P T Q G E V I P T MAdCAM-1 c
209 T P R S P G - - - - - - S T R T R R P E I S - - - - - - - - - - Q A G P T Q G E V I P T MAdCAM-1 d
188 T P R S P G - - - - - - S T R T R R P E I S - - - - - - - - - - Q A G P T Q G E V I P T MAdCAM-1 e

FIG.7C

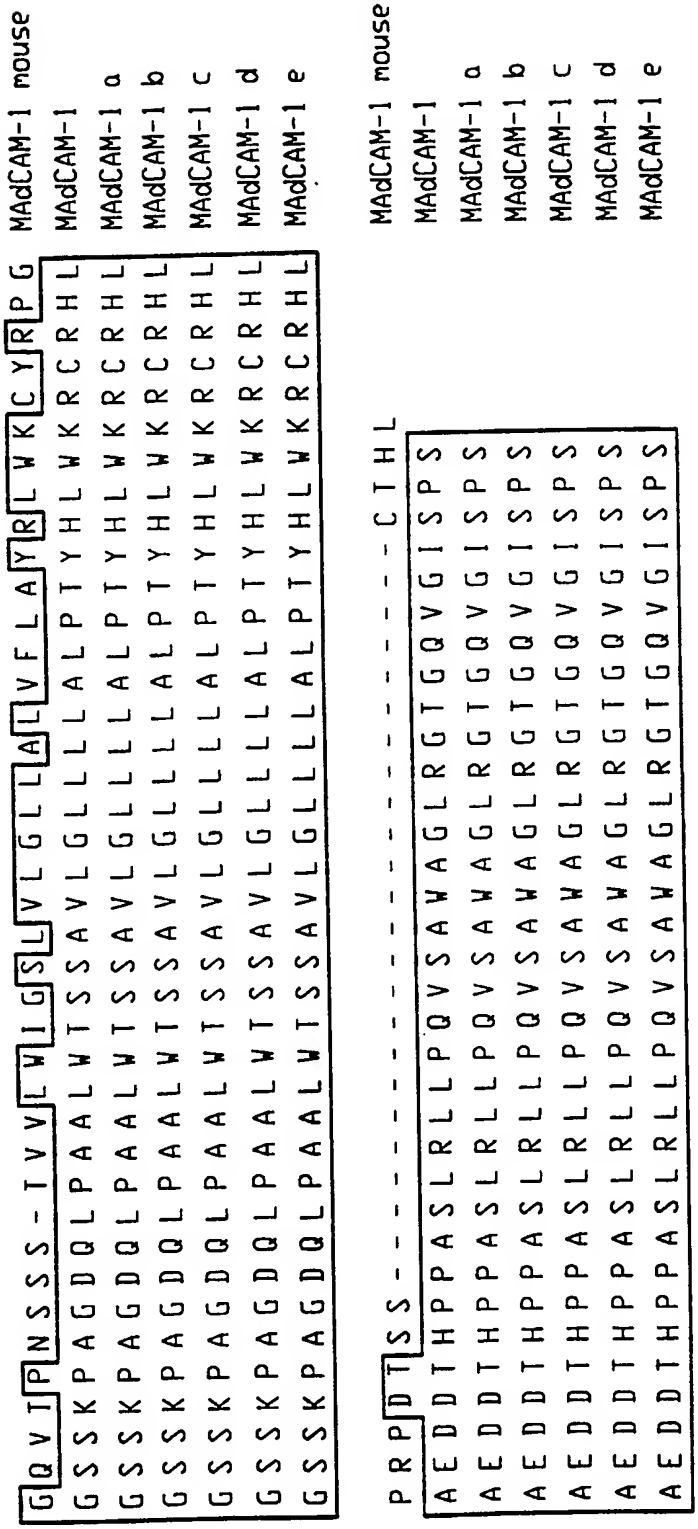

356 GQVTPNSSS-TVVLWIGSLVLGLLALVFLAYRLWKCYRPG MAdCAM-1 mouse
333 GSSKPAGDQLPAALWTSSAVLGLLLLALPTYHLWKRCRHL MAdCAM-1
309 GSSKPAGDQLPAALWTSSAVLGLLLLALPTYHLWKRCRHL MAdCAM-1 a
293 GSSKPAGDQLPAALWTSSAVLGLLLLALPTYHLWKRCRHL MAdCAM-1 b
190 GSSKPAGDQLPAALWTSSAVLGLLLLALPTYHLWKRCRHL MAdCAM-1 c
237 GSSKPAGDQLPAALWTSSAVLGLLLLALPTYHLWKRCRHL MAdCAM-1 d
216 GSSKPAGDQLPAALWTSSAVLGLLLLALPTYHLWKRCRHL MAdCAM-1 e
395 PRPDTSS----------CTHL MAdCAM-1 mouse
373 AEDDTHPPASLRLLPQVSAWAGLRGTGQVGISPS MAdCAM-1
349 AEDDTHPPASLRLLPQVSAWAGLRGTGQVGISPS MAdCAM-1 a
333 AEDDTHPPASLRLLPQVSAWAGLRGTGQVGISPS MAdCAM-1 b
230 AEDDTHPPASLRLLPQVSAWAGLRGTGQVGISPS MAdCAM-1 c
277 AEDDTHPPASLRLLPQVSAWAGLRGTGQVGISPS MAdCAM-1 d
256 AEDDTHPPASLRLLPQVSAWAGLRGTGQVGISPS MAdCAM-1 e

FIG.7D

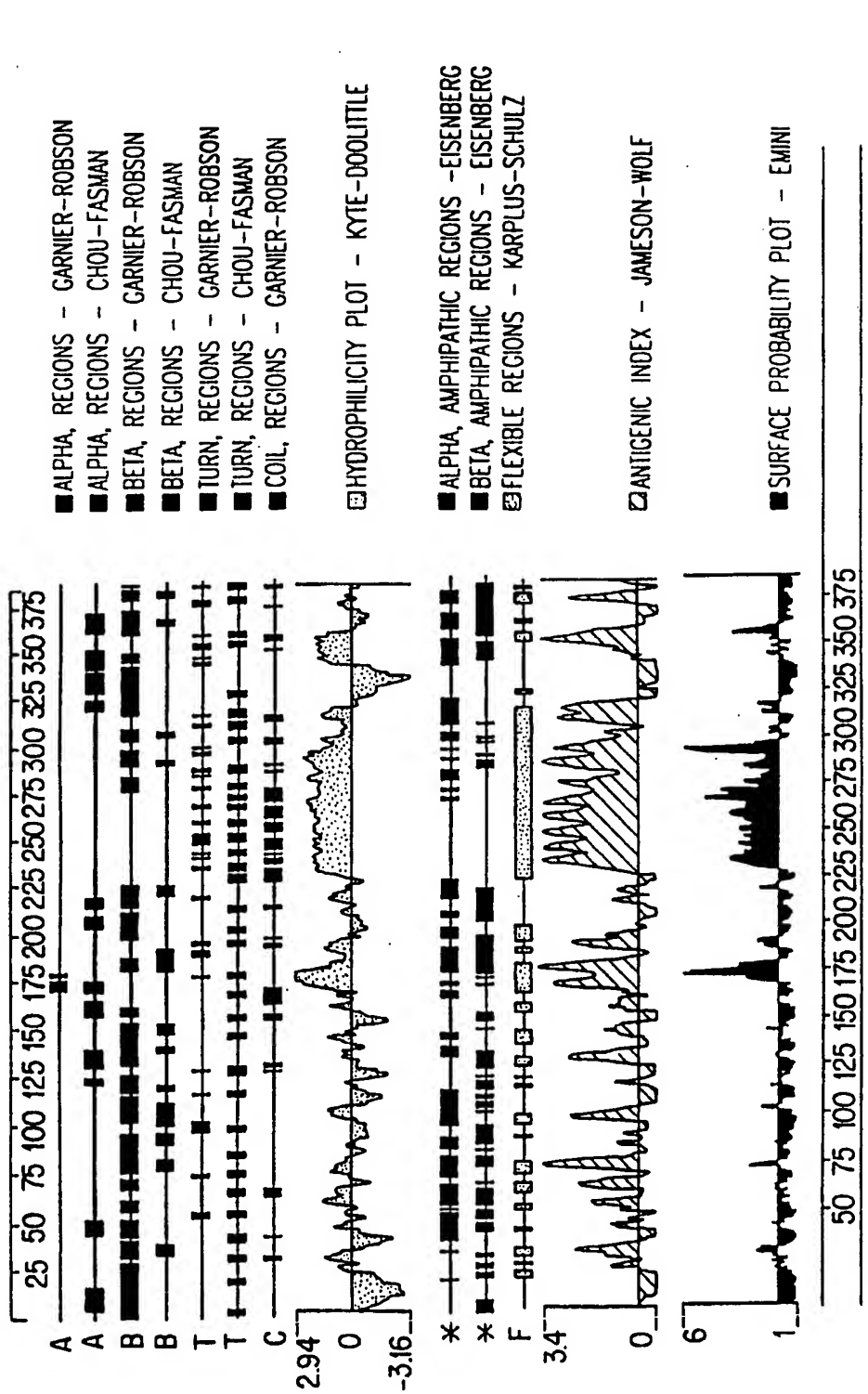

25 50 75 100 125 150 175 200 225 250 275 300 325 350 375
A
A
B
B
T
T
C
2.94
0
-3.16
ALPHA, REGIONS - GARNIER-ROBSON
ALPHA, REGIONS - CHOU-FASMAN
BETA, REGIONS - GARNIER-ROBSON
BETA, REGIONS - CHOU-FASMAN
TURN, REGIONS - GARNIER-ROBSON
TURN, REGIONS - CHOU-FASMAN
COIL, REGIONS - GARNIER-ROBSON
HYDROPHILICITY PLOT - KYTE-DOOLITTLE
*
*
F
3.4
0
ALPHA, AMPHIPATHIC REGIONS -EISENBERG
BETA, AMPHIPATHIC REGIONS - EISENBERG
FLEXIBLE REGIONS - KARPLUS-SCHULZ
ANTIGENIC INDEX - JAMESON-WOLF
6
1
SURFACE PROBABILITY PLOT - EMINI
50 75 100 125 150 175 200 225 250 275 300 325 350 375

FIG. 8

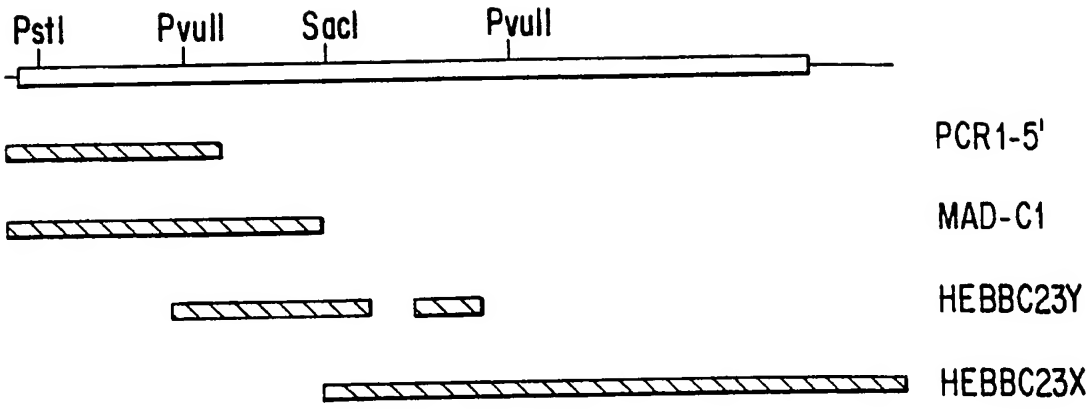


FIG. 9A

```
                                                    GGGACTGAGC  -1

ATGGATTTCGGACTGGCCCTCCTGCTGGCGGGGCTTCTGGGGCTCCTCCTCGGCCAGTCCCTCCAGGTGA   70
M  D  F  G  L  A  L  L  L  A  G  L  L  G  L  L  L  G  Q  S  L  Q  V  K    24

AGCCCCTGCAGGTGGAGCCCCCGGAGCCGGTGGTGGCCGTGGCCTTGGGCGCCTCGCGCCAGCTCACCTG  140
  P  L  Q  V  E  P  P  E  P  V  V  A  V  A  L  G  A  S  R  Q  L  T  Ⓒ    47

CCGCCTGGCCTGCGCGGACCGCGGGGCCTCGGTGCAGTGGCGGGGCCTGGACACCAGCCTGGGCGCGGTG  210
R  L  A  Ⓒ  A  D  R  G  A  S  V  Q  W  R  G  L  D  T  S  L  G  A  V      70

CAGTCGGACACGGGCCGCAGCGTCCTCACCGTGCGCAACGCCTCGCTGTCGGCGGCCGGGACCCGCGTGT  280
Q  S  D  T  G  R  S  V  L  T  V  R  N  A  S  L  S  A  A  G  T  R  V  Ⓒ    94

GCGTGGGCTCCTGCGGGGGCCGCACCTTCCAGCACACCGTGCAGCTCCTTGTGTACGCCTTCCCGGACCA  350
  V  G  S  Ⓒ  G  G  R  T  F  Q  H  T  V  Q  L  L  V  Y  A  F  P  D  Q    117

GCTGACCGTCTCCCCAGCAGCCCTGGTGCCTGGTGACCCGGAGGTGGCCTGTACGGCCCACAAAGTCACG  420
L  T  V  S  P  A  A  L  V  P  G  D  P  E  V  A  Ⓒ  T  A  H  K  V  T      140

CCCGTGGACCCCAACGCGCTCTCCTTCTCCCTGCTCGTCGGGGGCCAGGAACTGGAGGGGGCGCAAGCCC  490
P  V  D  P  N  A  L  S  F  S  L  L  V  G  G  Q  E  L  E  G  A  Q  A  L    164

TGGGCCCGGAGGTGCAGGAGGAGGAGGAGGAGCCCCAGGGGGACGAGGACGTGCTGTTCAGGGTGACAGA  560
  G  P  E  V  Q  E  E  E  E  E  P  Q  G  D  E  D  V  L  F  R  V  T  E    187

GCGCTGGCGGCTGCCGCCCCTGGGGACCCCTGTCCCGCCCGCCCTCTACTGCCAGGCCACGATGAGGCTG  630
R  W  R  L  P  P  L  G  T  P  V  P  P  A  L  Y  Ⓒ  Q  A  T  M  R  L      210

CCTGGCTTGGAGCTCAGCCACCGCCAGGCCATCCCCGTCCTGCACAGCCCGACCTCCCCGGAGCCTCCCG  700
P  G  L  E  L  S  H  R  Q  A  I  P  V  L  H  S  P  T  S  P  E  P  P  D    234

ACACCACCTCCCCGGAGTCTCCCGACACCACCTCCCCGGAGTCTCCCGACACCACCTCCCCGGAGCCTCC  770
  T  T  S  P  E  S  P  D  T  T  S  P  E  S  P  D  T  T  S  P  E  P  P    287

CGACACCACCTCCCCGGAGCCTCCCGACAAGACCTCCCCGGAGCCCGCCCCCAGCAGGGCTCCACACAC  840
D  T  T  S  P  E  P  P  D  K  T  S  P  E  P  A  P  Q  Q  G  S  T  H      280

ACCCCCAGGAGCCCAGGCTCCACCAGGACTCGCCGCCCTGAGATCTCCCAGGCTGGGCCCACGCAGGGAG  910
T  P  R  S  P  G  S  T  R  T  R  R  P  E  I  S  Q  A  G  P  T  Q  G  E    304

AAGTGATCCCAACAGGCTCGTCCAAACCTGCGGGTGACCAGCTGCCCGCGGCTCTGTGGACCAGCAGTGC  980
  V  I  P  T  G  S  S  K  P  A  G  D  Q  L  P  A  A  L  W  T  S  S  A    327

GGTGCTGGGACTGCTGCTCCTGGCCTTGCCCACGTATCACCTCTGGAAACGCTGCCGGCACCTGGCTGAG 1050
V  L  G  L  L  L  L  A  L  P  T  Y  H  L  W  K  R  C  R  H  L  A  E      350

GACGACACCCACCCACCAGCTTCTCTGAGGCTTCTGCCCCAGGTGTCGGCCTGGGCTGGGTTAAGCGCGA 1120
D  D  T  H  P  P  A  S  L  R  L  L  P  Q  V  S  A  W  A  G  L  R  G  T  374
```

FIG.9B

```
CCGGCCAGGTCGGGATCAGCCCCTCCTGAGTGGCCAGCCTTTCCCCCTGTGAAAGCAAAATAGCTTGGAC 1190
  G  Q  V  G  I  S  P  S                                                382

CCCTTCAAGTTGAGAACTGGTCAGGGCAAACCTGCCTCCCATTCTACTCAAAGTCATCCCTCTGTTCACA 1260

GAGATGGATGCATGTTCTGATTGCCTCTTTGGAGAAGCTCATCAGAAACTCAAAAGAAGGCCACTGTTTG 1330

TCTCACCTACCCATGACCTGAAGCCCCTCCCTGAGTGGTCCCCACCTTTCTGGACGGAACCACGTACTTT 1400

TTACATACATTGATTCATGTCTCACGTCTCCCTAAAAATGCGTAAGACCAAGCTGTGCCCTGACCACCCT 1470

GGGCCCCTGTCGTCAGGACCTCCTGAGGCTTTGGCAAATAAACCTCCTAAAATGATAAAAAAAAAA     1536
```

FIG.9C

SPTSPEPP
DTTSPESP
DTTSPESP
DTTSPEPP
DTTSPEPP
DKTSPEPA

FIG.10A

```
228  TSPEPPDTTSPESPDTTSPESPDTTSPEPPDTTSPEPPDKTSPEP   272  MAdCAM-1
     T+P PP TT+   P TT+P  P TT+  PP TT+P PP  T+  P
 78  TTPSPPTTTTTTPPPTTTPSPPITTTTTPPPTTTPSPPISTTTTP   122  Muc-2
```

FIG.10B

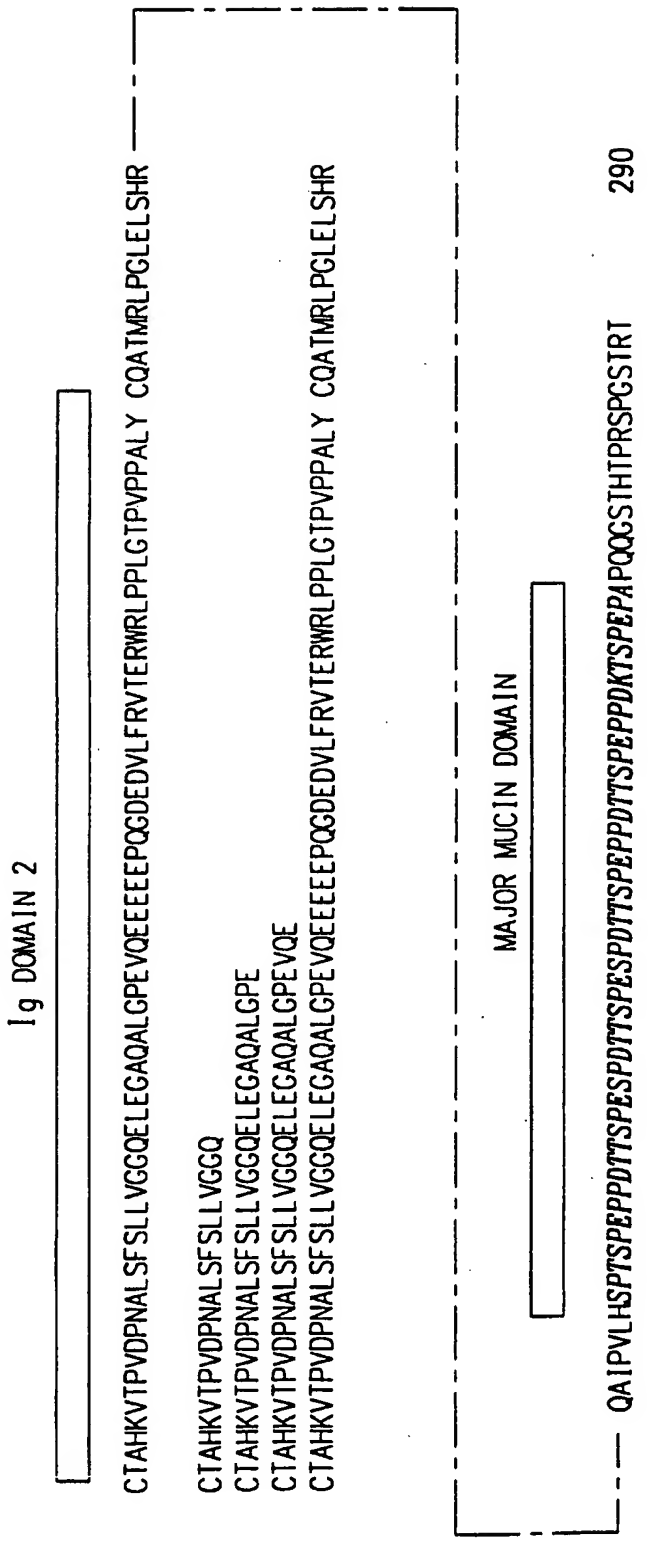

Ig DOMAIN 2
CTAHKVTPVDPNALSFSLLVGGQELEGAQALGPEVQEEEEEPQGDEDVLFRVTERWRLPPLGTPVPPALY CQATMRLPGLELSHR
CTAHKVTPVDPNALSFSLLVGGQ
CTAHKVTPVDPNALSFSLLVGGQELEGAQALGPE
CTAHKVTPVDPNALSFSLLVGGQELEGAQALGPEVQE
CTAHKVTPVDPNALSFSLLVGGQELEGAQALGPEVQEEEEEPQGDEDVLFRVTERWRLPPLGTPVPPALY CQATMRLPGLELSHR
MAJOR MUCIN DOMAIN
QAIPVLHSPTSPEPPDTTSPESPDTTSPEPPDTTSPEPPDKTSPEPAPQQGSTHTPRSPGSTRT 290
QGSTHTPRSPGSTRT
SPDTTSPEPPDTTSPEPPDKTSPEPAPQQGSTHTPRSPGSTRT
PPDKTSPEPAPQQGSTHTPRSPGSTRT
QAIPVLHSPTSPB
PPDTTSPEPPDKTSPEPAPQQGSTHTPRSPGSTRT

FIG.11A

	exon	5' splice donor	3' splice acceptor	exon
7	CGGGGGCCAG	GAACTGGAGG	CGCCCCCCAG	CAGGGCTCCA
5	GGGCCCGGAG	GTGCAGGAGG	CTCCCCGGAG	TCTCCCGACA
3	GGTGCAGGAG	GAGGAGGAGG	CTCCCCGGAG	CCTCCCGACA
y	CTCCCCGGAG	CCTCCCGACA	CTCCCCGGAG	CCTCCCGACA

FIG.11B

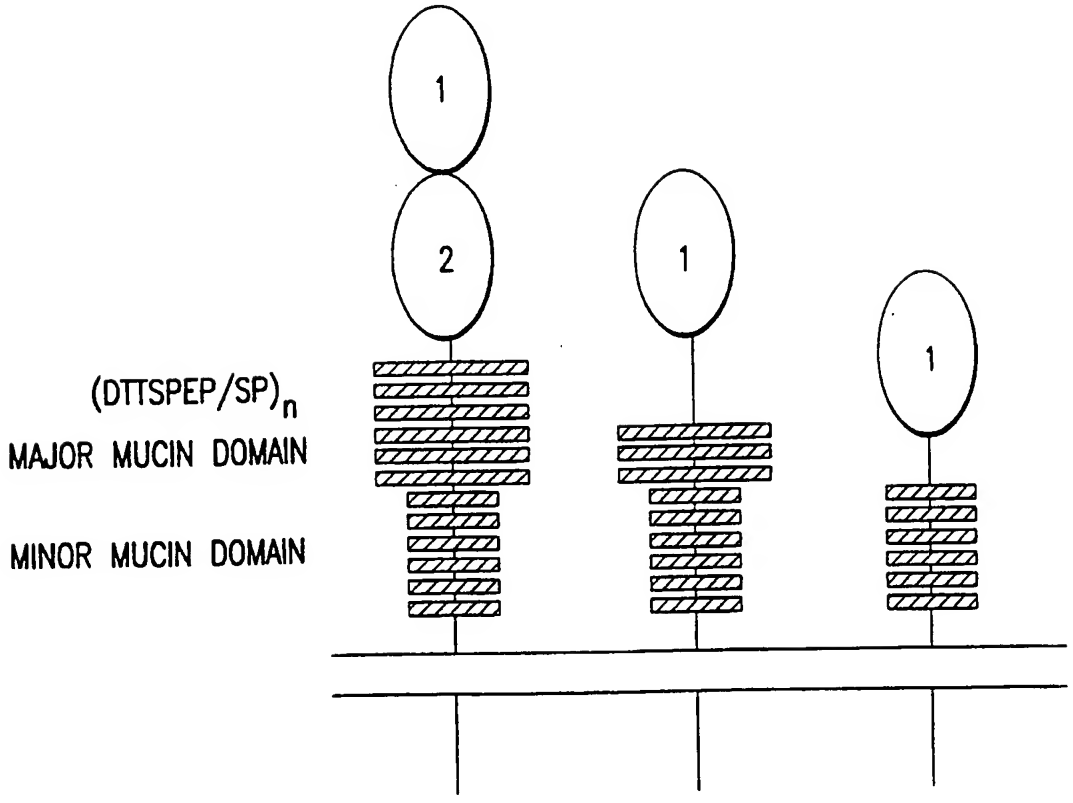


FIG.12

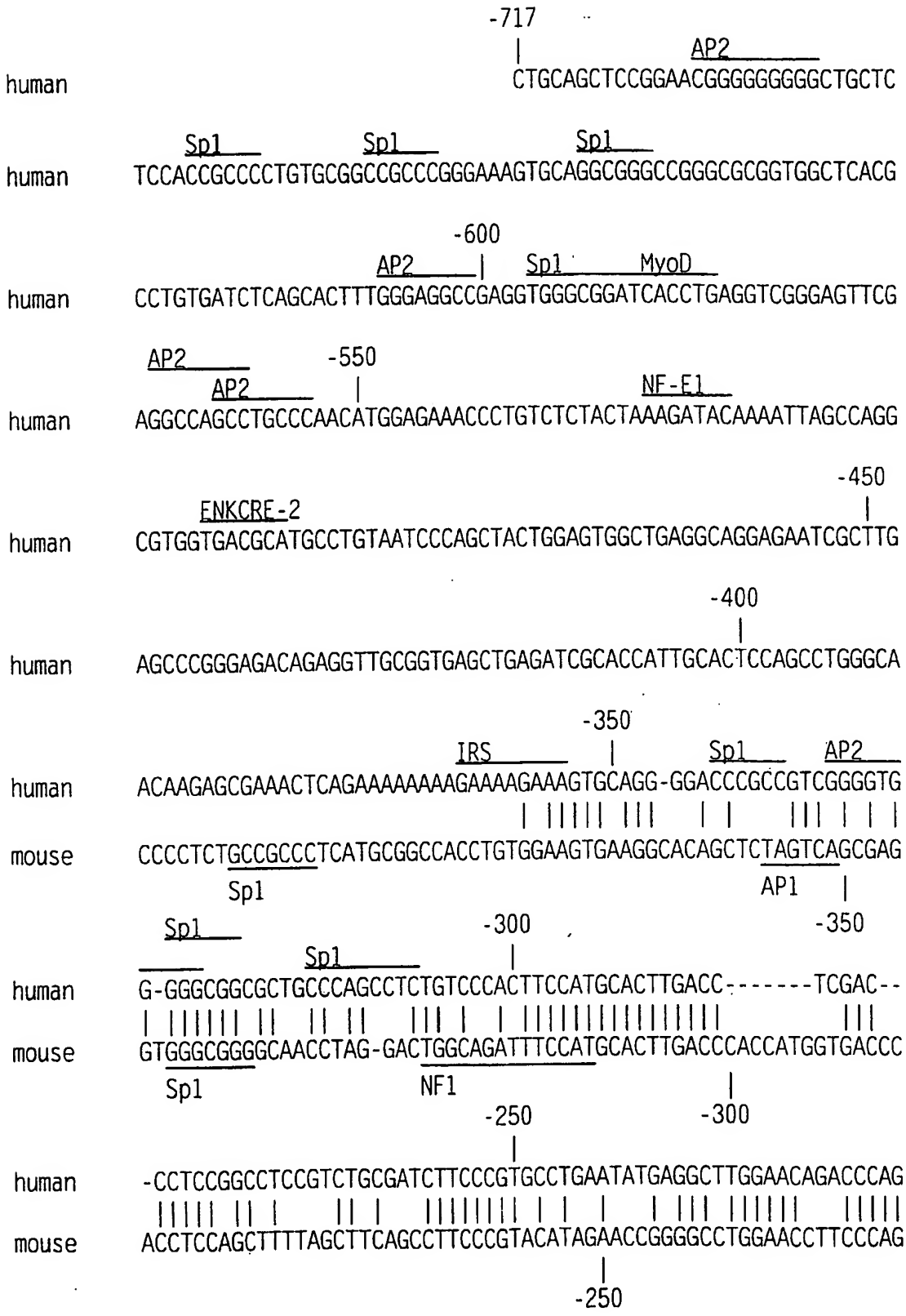

-717
AP2
human CTGCAGCTCCGGAACGGGGGGGGGCTGCTC
Sp1 Sp1 Sp1
human TCCACCGCCCCTGTGCGGCCGCCCGGGAAAGTGCAGGCGGGCCGGGCGCGGTGGCTCACG
-600
AP2 Sp1 MyoD
human CCTGTGATCTCAGCACTTTGGGAGGCCGAGGTGGGCGGATCACCTGAGGTCGGGAGTTCG
AP2 -550
AP2 NF-E1
human AGGCCAGCCTGCCCAACATGGAGAAACCCTGTCTCTACTAAAGATACAAAATTAGCCAGG
-450
ENKCRE-2
human CGTGGTGACGCATGCCTGTAATCCCAGCTACTGGAGTGGCTGAGGCAGGAGAATCGCTTG
-400
human AGCCCGGGAGACAGAGGTTGCGGTGAGCTGAGATCGCACCATTGCACTCCAGCCTGGGCA
-350
IRS Sp1 AP2
human ACAAGAGCGAAACTCAGAAAAAAAAGAAAAGAAAGTGCAGG-GGACCCGCCGTCGGGGTG
mouse CCCCTCTGCCGCCCTCATGCGGCCACCTGTGGAAGTGAAGGCACAGCTCTAGTCAGCGAG
Sp1 AP1
Sp1 -300 -350
Sp1
human G-GGGCGGCGCTGCCCAGCCTCTGTCCCACTTCCATGCACTTGACC-------TCGAC--
mouse GTGGGCGGGGCAACCTAG-GACTGGCAGATTTCCATGCACTTGACCCACCATGGTGACCC
Sp1 NF1
-250 -300
human -CCTCCGGCCTCCGTCTGCGATCTTCCCGTGCCTGAATATGAGGCTTGGAACAGACCCAG
mouse ACCTCCAGCTTTTAGCTTCAGCCTTCCCGTACATAGAACCGGGGCCTGGAACCTTCCCAG
-250

FIG.13A

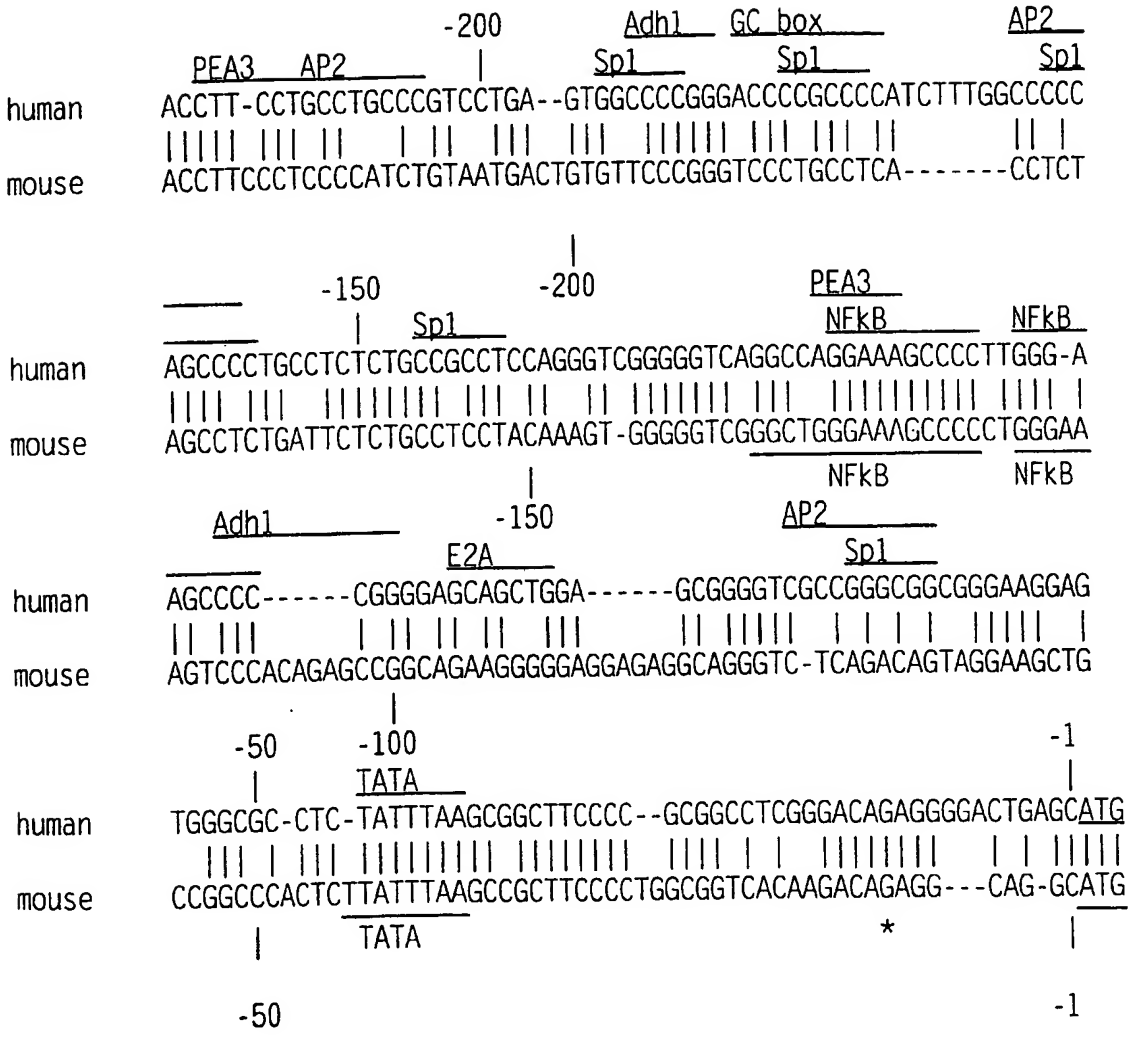

-200
Adh1
GC box
AP2
PEA3
AP2
Sp1
Sp1
Sp1
human
ACCTT-CCTGCCTGCCCGTCCTGA--GTGGCCCCGGGACCCCGCCCCATCTTTGGCCCCC
mouse
ACCTTCCCTCCCCATCTGTAATGACTGTGTTCCCGGGTCCCTGCCTCA-------CCTCT
-150
-200
PEA3
Sp1
NFkB
NFkB
human
AGCCCCTGCCTCTCTGCCGCCTCCAGGGTCGGGGGTCAGGCCAGGAAAGCCCCTTGGG-A
mouse
AGCCTCTGATTCTCTGCCTCCTACAAAGT-GGGGGTCGGGCTGGGAAAGCCCCCTGGGAA
NFkB
NFkB
Adh1
-150
AP2
E2A
Sp1
human
AGCCCC------CGGGGAGCAGCTGGA------GCGGGGTCGCCGGGCGGCGGGAAGGAG
mouse
AGTCCCACAGAGCCGGCAGAAGGGGGAGGAGAGGCAGGGTC-TCAGACAGTAGGAAGCTG
-50
-100
-1
TATA
human
TGGGCGC-CTC-TATTTAAGCGGCTTCCCC--GCGGCCTCGGGACAGAGGGGACTGAGCATG
mouse
CCGGCCCACTCTTATTTAAGCCGCTTCCCCTGGCGGTCACAAGACAGAGG---CAG-GCATG
TATA
*
-50
-1

FIG.13B

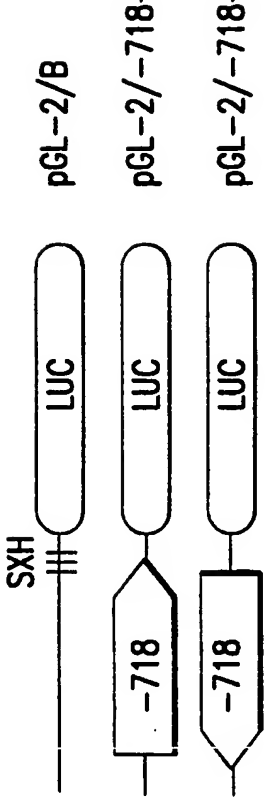


FIG. 14A

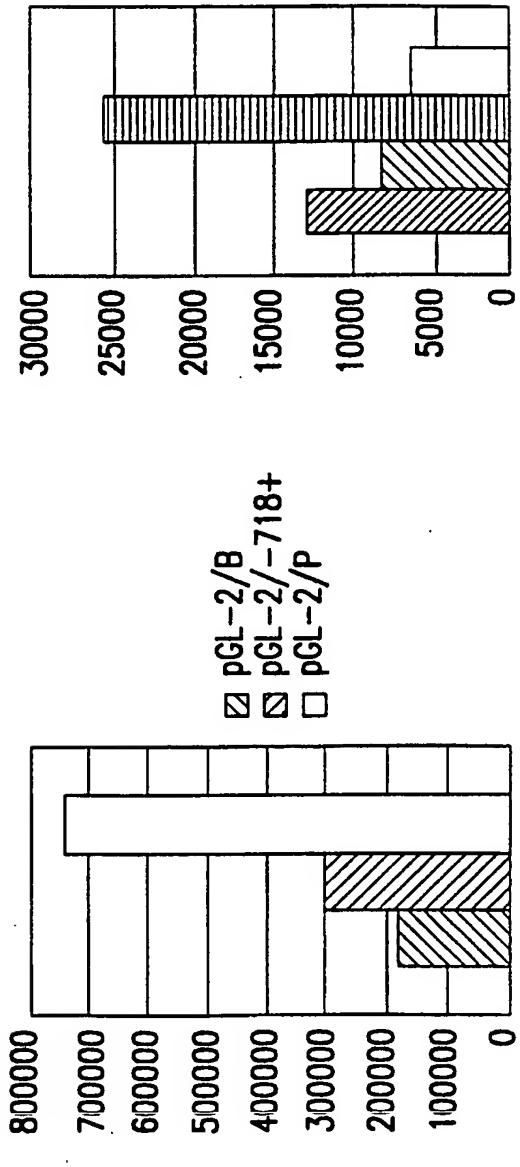


FIG. 14B

FIG. 14C

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/17549

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12N 1/21, 5/10, 15/12, 15/64, 15/66, 15/70, 15/79; C07K 14/705
US CL :536/23.5, 24.1; 435/69.1, 240.2, 252.3, 320.1; 530/395
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 536/23.5, 24.1; 435/69.1, 240.2, 252.3, 320.1; 530/395

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS/USPAT; STN/Medline, HCAPlus
search terms: MAdCAM#, mucosal vascular addressin cell adhesion molecule#, mad cam#

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X ---- Y	SHYJAN, A.M. et al. Human Mucosal Addressin Cell Adhesion Molecule-1 (MAdCAM-1) Demonstrates Structural and Functional Similarities to the Alpha 4 Beta 7-Integrin Binding Domains of Murine MAdCAM-1, but Extreme Divergence of Mucin-Like Sequences. Journal of Immunology. 15 April 1996, Volume 156, Number 8, pages 2851-2857, especially pages 2852-2854.	1-17 ---- 1-17, 21-23
Y	SAMPAIO, S.O. et al. Organization, Regulatory Sequences, and Alternatively Spliced Transcripts of the Mucosal Addressin Cell Adhesion Molecule-1 (MAdCAM-1) Gene. Journal of Immunology. 01 September 1995, Volume 155, Number 5, pages 2477-2486, especially pages 2479-2482.	1-17, 21-23

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 02 FEBRUARY 1997	Date of mailing of the international search report 21 FEB 1997
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer KAREN E. BROWN Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/17549

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCHIFFER, S.G. et al. An Alternately Spliced mRNA Encoding Functional Domains of Murine MAdCAM-1. Biochemical and Biophysical Research Communications. 02 November 1995, Volume 216, Number 1, pages 170-176, especially page 173.	1-17
Y	TAKEUCHI, M. et al. Induction of the Gene Encoding Mucosal Vascular Addressin Cell Adhesion Molecule 1 by Tumor Necrosis Factor Alpha is Mediated by NF-Kappa B Proteins. Proceedings of the National Academy of Science USA, April 1995, Volume 92, pages 3561-3565, especially pages 3562-3563.	21-23

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/17549

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 1-17 and 21-23 (in part)
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

Because a computer-readable copy of the sequence listing was not available, all of the claims were unsearchable to the extent that no meaningful search of the sequences per se can be carried out by this Authority. However, the subject matter of the claims has been searched to the extent possible with reference to the balance of the description.

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-17 and 21-23 (in part)

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

CORRECTED VERSION*

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 1/21, 5/10, 15/12, 15/64, 15/66, 15/70, 15/79, C07K 14/705

A1

(11) International Publication Number: WO 98/20110

(43) International Publication Date: 14 May 1998 (14.05.98)

(21) International Application Number: PCT/US96/17549

(22) International Filing Date: 1 November 1996 (01.11.96)

(71) Applicants *(for all designated States except US)*: HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850 (US). THE UNIVERSITY OF AUCKLAND [NZ/NZ]; 85 Park Road, Grafton, Auckland 1001 (NZ).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: NI, Jian [CN/US]; 5502 Manorfield Road, Rockville, MD 20853 (US). GREENE, John, M. [US/US]; 872 Diamond Drive, Gaithersburg, MD 20878 (US). KRISSANSEN, Geoffrey, W. [NZ/NZ]; 157B Grand Drive, St. Johns, Auckland 1001 (NZ). LEUNG, Euphemia, Yee, Fun [US/NZ]; 10 Martin Avenue, Remuera, Auckland 1001 (NZ). RUBEN, Steven, M. [US/US]; 18528 Heritage Hills Drive, Olney, MD 20832 (US).

(74) Agents: GOLDSTEIN, Jorge, A. et al.; Sterne, Kessler, Goldstein & Fox P.L.L.C., Suite 600, 1100 New York Avenue, N.W., Washington, DC 20005–3934 (US).

(81) Designated States: CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
With an indication in relation to a deposited microorganism furnished under Rule 13bis separately from the description.
Date of receipt by the International Bureau:
10 October 1996 (10.10.1996)

(54) Title: HUMAN MUCOSAL ADDRESSIN CELL ADHESION MOLECULE–1 (MAdCAM–1) AND SPLICE VARIANTS THEREOF

(57) Abstract

The present invention relates to novel MAdCAM–1 proteins designated herein as MAdCAM–1(a–e), which are cell adhesion molecules. In particular, isolated nucleic acid molecules are provided encoding the human MAdCAM–1(a–e) proteins. MAdCAM–1(a–e) polypeptides are also provided as are vectors, host cells and recombinant methods for producing the same. The invention further relates to screening methods for identifying agonists and antagonists of MAdCAM–1(a–e) activity. Also provided are diagnostic methods for detecting cancer or a pathological inflammatory condition, and therapeutic methods for treating an individual in need of a reduction in the activity of any of MAdCAM–1(a–e). In another aspect, the invention provides isolated genomic DNA molecules comprising the 5 exons which comprise the genes which encode any of MAdCAM–1(a–e), as well as the 5' flanking region which includes the promoter for these genes. In another aspect, the invention relates to a method of screening compounds for the ability to regulate expression of any of MAdCAM–1(a–e) from their promoter. The invention also relates to a method of selectively expressing genes on gut endothelia.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

This Page Blank (uspto)